Design of Objective Quality Measures for Time-Scale Modification of Audio

GRIFFITH UNIVERSITY

School of Engineering and Built Environment

Signal Processing Laboratory

Design of Objective Quality Measures for Time-Scale Modification of Audio

Timothy Roberts

BEng (Hons) (Elec.Comp.), Griffith University, 2016

BMus (Tech), Griffith University, 2009

GUID: s2599923

September 2020

Dissertation submitted in fulfilment
of the requirements of the degree of

Doctor of Philosophy

Principal Supervisor: Professor Kuldip K. Paliwal

Associate Supervisor: Doctor Andrew Busch

Abstract

Time-Scale Modification (TSM) is a well-researched field and allows for time-domain manipulation of a signal without modifying the pitch or timbre. Many TSM methods have been presented, however quantitative results on the quality of these methods are rare, with most methods reporting informal listening tests. This is likely due to the time-commitment and cost of subjective testing. Additionally, an objective measure of quality has not yet been developed that is suitable for time-scaled signals. This dissertation describes the design of effective objective measures of quality for TSM.

TSM methods are, generally, single channel algorithms that give poor results when applied to multi-channel signals, as the phase relationship between channels must be maintained. This dissertation proposes a method and additional variant for maintaining the phase relationship between channels and retaining the presence in the centre of the stereo signal. The method involves pre- and post-processing the signal, with the variant processing each frame for real-time suitability. Sum and difference transformations of the stereo signal are used for TSM and result in a large improvement in stereo phase coherence, consequently maintaining the stereo field. The proposed method produces a high-quality stereo output and greatly improves quality over the independent

channel processing method. It also allows for simple implementation around all existing TSM frameworks.

A modification to the Epoch-Synchronous Overlap-Add (ESOLA) TSM algorithm is proposed in this dissertation. The proposed method, Fuzzy Epoch-Synchronous Overlap-Add, improves on the previous ESOLA method through cross-correlation of time-smeared epochs before overlap-adding. This reduces distortion and artefacts while the speaker's fundamental frequency is stable, as well as reducing artefacts during pitch modulation. The proposed method is tested against well-known TSM algorithms. It is preferred over ESOLA and gives similar performance to other TSM algorithms for voice signals. It is also shown that this algorithm can work effectively with solo instrument signals containing strong fundamental frequencies.

No effective objective measure of quality for TSM exists. This dissertation details the creation, subjective evaluation and analysis of a dataset, for use in the development of an objective measure of quality for TSM. Comprising two parts, the training subset contains 88 source files processed using six TSM methods at 10 time-scales, while the testing subset contains 20 source files processed using three additional methods at four time-scales. The source material contains speech, solo harmonic and percussive instruments, sound effects and a range of music genres. 42,529 ratings were collected from 633 sessions using laboratory and remote collection methods. Analysis of results shows no correlation between age and quality of rating; equivalence between expert and non-expert listeners; negligible differences between participants with and without hearing issues; and negligible differences between testing modalities. Comparison of published objective measures and subjective scores shows the objective measures to be poor indicators of subjective quality.

Initial results for a retrained objective measure of quality are presented with results approaching average loss and correlation values of subjective sessions.

An objective measure of quality for time-scaled audio is proposed that makes use of the previously developed dataset and improves on reported results. The measure uses hand-crafted features and a fully connected network to predict subjective mean opinion scores. Basic and Advanced Perceptual Evaluation of Audio Quality features are used in addition to nine features specific to TSM artefacts. Six methods of alignment are explored, with interpolation of the reference magnitude spectrum to the length of the test magnitude spectrum giving the best performance. The proposed measure achieves an average Root Mean Squared Error (RMSE) of 0.490 and a mean Pearson Correlation Coefficient (PCC) of 0.864, equivalent to 97th and 82nd percentiles of subjective sessions respectively. The proposed measure is used to evaluate TSM algorithms, finding that Elastique gives the highest objective quality for solo instrument and voice signals, while the Identity Phase-Locking Phase Vocoder gives the highest objective quality for music signals and the best overall quality.

Two single-ended objective quality measures for time-scaled audio are also proposed. These measure do not require a reference signal, nor alignment. Data driven features are created by either a convolutional neural network (CNN) or a bidirectional gated recurrent unit (BGRU) network, and are fed to a fully-connected network to predict subjective mean opinion scores. The proposed CNN and BGRU measures achieve an average RMSE of 0.608 and 0.576, and a mean PCC of 0.771 and 0.794, respectively. The proposed measures are used to evaluate TSM algorithms, and comparisons are provided for 16 TSM implementations.

A literature review is included with required background knowledge. It includes the fundamentals of sound perception, sound capture, digital signal processing, time-scale modification methods used within research, and subjective and objective measures of quality.

Full implementation of all proposed methods and measures can be found at github.com/zygurt/TSM, while the labelled dataset is available at http://ieee-dataport.org/1987.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Acknowledgements

I would first like to express unending thanks and gratitude to Kelsie, for your love, care, support and loyalty. Thank you for keeping me grounded, reminding me of the bigger picture and sacrificing things that are important to you, so that I could do this. Thank you, Mum, for continuing to ask questions as my research changed over time, and Dad, for your ears that are now well trained. I greatly appreciate you both. To my immediate and extended family, thank you for asking questions and allowing me to gain a greater understanding of, and increasing my ability to explain, my research. To the Signal Processing Lab and friends: Aaron, Jack, Aadel, Jaswinder, Sujan, Jaspreet, Utkarsh, Peyush, Nick, Jordan, Anil, Sisi, James and Rhys, thank you for the conversation, wider field discussion, machine learning assistance and time volunteered for listening tests. I would also like to show my immense gratitude to everyone who assisted in the subjective testing, this couldn't have happened without your help. To Dr. Andrew Busch, thank you for your guidance before and through the PhD. Finally, I would like to thank my principle supervisor, Prof. Kuldip Paliwal, for allowing and trusting me to explore this field.

Galatians 1:5

Contents

Abstract iii

Statement of Originality vii

Acknowledgements ix

Contents xi

List of Figures xix

List of Tables xxix

Table of Abbreviations xxxi

Table of Symbols xxxvii

I Introduction 1

1 Introduction 3

1.1 Background 3

1.2 Dissertation Organisation 5

1.3 Contributions 8

1.4 Publications 9

1.4.1 Publications Resulting from Dissertation Research 9

1.4.2 Additional Publications 10

1.5 Ethical Clearance for Experimentation 11

II Literature Review **13**

2 Fundamentals **15**

2.1 Sound Perception 15

2.2 Sound Capture 19

2.2.1 Analogue to Digital Conversion 20

2.3 Digital Signal Processing 22

2.3.1 Time Domain 22

2.3.2 Spectral Domain 25

2.3.3 Analysis Modification Synthesis Framework 26

3 Time-Scale Modification **29**

3.1 Historical Methods 30

3.2 Time Domain Methods 31

3.2.1 Overlap-Add 31

3.2.2 Synchronised Overlap-Add 32

3.2.3 Waveform Similarity Overlap-Add 33
3.2.4 Epoch Synchronous Overlap-Add 35
3.3 Frequency Domain Methods 37
3.3.1 Phase Vocoder 38
3.3.2 Phase-Locking Phase Vocoders 41
3.3.3 Multi-Resolution Peak Picking and Sinusoidal Trajectory Heuristics 44
3.3.4 Fuzzy Classification of Spectral Bins 47
3.3.5 Mel-Scale Sub-band Modelling 50
3.4 Source Separation Methods 51
3.4.1 Harmonic Percussive Time-Scale Modification . . . 51
3.4.2 Non-Negative Matrix Factorisation Time-Scale Modification 54
3.5 Other Methods 54
3.5.1 Published Methods 54
3.5.2 Commercial Methods 55
3.6 Additional Information 55
3.6.1 Multi-channel Considerations 55
3.6.2 Matlab TSM library 58

4 Measures of Quality **59**

4.1 Introduction 59
4.2 Grading Scales 60

4.3 Subjective Measures 61
4.3.1 General Methods 61
4.3.2 MUSHRA 62
4.3.3 Subjective Testing of Time-Scale Modification . . . 63
4.4 Objective Measures 64
4.4.1 Traditional Time-Scale Modification Measures . . 64
4.4.2 Signal to Noise Ratio 66
4.4.3 Perceptual Evaluation of Speech Quality 67
4.4.4 Composite Speech Quality Measures 69
4.4.5 PEAQ 70
4.4.6 Related Objective Measures of Quality 74
4.4.7 Figures of Merit 75

III Research **77**

5 Time-Scale Modification Improvements **79**

5.1 Stereo Time-Scale Modification 79
5.1.1 Introduction 79
5.1.2 Method 81
5.1.3 Testing 83
5.1.4 Results 88
5.1.5 Conclusion 94

5.2 Fuzzy Epoch-Synchronous Overlap-Add 95

5.2.1 Introduction 95

5.2.2 Background 95

5.2.3 Method 96

5.2.4 Testing Methodology 98

5.2.5 Results 98

5.2.6 Conclusion 103

6 A Time-Scale Modification Dataset with Subjective Quality Labels 105

6.1 Introduction 105

6.2 Algorithms and Quality Evaluation 108

6.3 Dataset Description 112

6.4 Subjective Testing 114

6.4.1 Results 117

6.5 Towards an Objective Measure of Quality 130

6.6 Conclusion 132

7 An Objective Measure of Quality for Time-Scale Modification of Audio 133

7.1 Introduction 133

7.2 Method 136

7.2.1 Changes to PEAQ 136

7.2.2 Additional Features 140
7.2.3 Feature Preparation 148
7.2.4 Network Structure 149
7.3 Results 150
7.3.1 Feature Results 150
7.3.2 Network Performance 153
7.3.3 TSM Algorithm Evaluation 157
7.4 Availability 159
7.5 Future Research 159
7.6 Conclusion 161

8 Deep Learning-Based Single-Ended Objective Quality Measures for Time-Scale Modified Audio 163

8.1 Introduction 163
8.2 Method 168
8.3 Results 173
8.3.1 Network Performance 173
8.3.2 TSM Algorithm Evaluation 179
8.4 Availability 187
8.5 Future Research 187
8.6 Conclusion 187

IV Conclusion **189**

9 Dissertation summary: Conclusions and Future Research **191**

9.1 Chapter 5.1: Stereo Time-Scale Modification 191

9.2 Chapter 5.2: Fuzzy Epoch-Synchronous Overlap-Add . . . 192

9.3 Chapter 6: A Time-Scale Modification Dataset with Subjective Quality Labels 192

9.4 Chapter 7: An Objective Measure of Quality for Time-Scale Modification of Audio 193

9.5 Chapter 8: Deep Learning-Based Single-Ended Objective Quality Measures for Time-Scale Modified Audio 194

9.6 Future Research 194

V Appendices **197**

TSM Dataset Reference File Listings **199**

A Training Subset 199

B Testing Subset 205

C Unused Reference File Subset 207

Bibliography **211**

List of Figures

2.1 The external and internal ear. (Flanagan, 1972) *Reprinted with permission.* 16

2.2 Fletcher-Munson equal-loudness curve. (Fletcher and Munson, 1933) *Reprinted with permission.* 17

2.3 Comparison between Uniform Frequency Scale and Frequency Warped Mel and Bark Scales. 19

2.4 Analogue to Digital Converter Block Diagram. 21

2.5 Waveform for a male utterance of "I am sitting in a room." 22

2.6 Framing an audio signal. 23

2.7 Shape of common window functions. 24

2.8 Spectrogram for male uttering "I am sitting in a room." . 27

2.9 Analysis Modification Synthesis Algorithm 28

3.1 [Colour Online] WSOLA Algorithm. (Driedger and Muller, 2016) *Reprinted under CC BY.* 34

3.2 Epochs within male speech calculated using the Zero Frequency Resonator method. 35

3.3 Epochs within a solo flute recording calculated using the Zero Frequency Resonator method. 36

3.4 [Colour Online] Sinusoidal Trajectories from Karrer et al. (2006). *Reprinted with permission.* 46

3.5 [Colour Online] Harmonic Percussive separation using binary masking. (Driedger et al., 2014) *Reprinted with permission. ©2013 IEEE.* 52

3.6 Bonada phase modification for preserving the stereo field. (Bonada, 2002). *Reprinted under CC BY-NC-ND 3.0 ES license.* 56

3.7 Altoe block diagram for preserving the stereo field. (Altoe, 2012). *Reprinted with permission.* 57

4.1 PESQ Block Diagram. (Loizou, 2013). *Reprinted with permission.* 67

4.2 PEAQ Block Diagram. (ITU-T, 2001a). *Reprinted with permission.* 71

5.1 Block diagram for the proposed file method. Preprocessing transforms the signal to the sum and difference representation, while post-processing transforms the scaled signal back to left and right. 81

5.2 Stereo features for white noise cross-fading in the sum and difference representation and a sine tone panning right to left. 85

5.3 Mean Subjective Preference score for Altoe, Bonada, Proposed and Naive Stereo TSM methods. Error bars show ±1 standard deviation from the mean. 89

5.4 Mean SPC dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better. 90

5.5 Mean stereo balance dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better. 90

5.6 [Colour Online] Mean SPC for multiple TSM methods at multiple TSM ratios. Dotted lines segment files. Timescale ratio increases left to right within each segment. . . 92

5.7 [Colour Online] Mean balance for multiple TSM methods at multiple TSM ratios. Dotted lines segment files. Timescale ratio increases left to right within each segment. . . 93

5.8 Epoch Synchronisation of the ESOLA method. 96

5.9 Epoch Synchronisation of the proposed FESOLA method. 99

5.10 ESOLA and FESOLA spectrogram for the *Child 1* file slowed to 53.83%. Distortion around the fundamental and partials are visible for the ESOLA method. Excerpt begins at 4.5 seconds through each file. 100

5.11 Mean preference comparison for ESOLA and FESOLA. . 101

5.12 Comparison of mean opinion scores averaged across all voice processed files. MOS of 1-5 is Bad-Excellent. 101

5.13 Comparison of mean opinion scores averaged across all solo processed files. MOS of 1-5 is Bad-Excellent. 102

5.14 Comparison of mean opinion scores averaged across all music processed files. MOS of 1-5 is Bad-Excellent. 102

5.15 Box plots of Mean Opinion Scores for FESOLA across all source files. MOS of 1-5 is Bad-Excellent. 103

5.16 Mean MOS for all Voice signals. MOS of 1-5 is Bad-Excellent. 104

6.1 Bespoke MATLAB graphical user interface used for laboratory sessions. 116

6.2 Web Audio Evaluation Tool user interface used for remote testing. Shown with two file pairs. 116

6.3 [Colour Online] Distribution of PCC and RMSE for all sessions before normalisation and outlier removal. Blue plus symbols mark PCC outliers, while red crosses mark RMSE outliers. 119

6.4 [Colour Online] Distribution of PCC and RMSE for each session before normalisation. Horizontal and vertical lines denote means. 119

6.5 [Colour Online] 2D Histogram of normalised responses, ordered by ascending MOS (red line). 121

6.6 Overall means for each method at each time-scale for all evaluated files. 122

6.7 Mean MOS for each method at each time-scale for musical source material. 124

6.8 Mean MOS for each method at each time-scale for solo instrument source material. 124

6.9 Mean MOS for each method at each time-scale for Voice source material. 125

6.10 [Colour Online] MOS standard deviation against the number of responses for that file. 126

6.11 TOST (1-α)100% CI for equivalence of participant experience for $\alpha = 0.05$. Equivalence interval of $\pm 5\%$ of expert participant means. 127

6.12 TOST (1-α)100% CI for equivalence of means of participants with and without hearing issues for $\alpha = 0.05$. Equivalence interval of $\pm 5\%$ of mean for participants without hearing issues. 128

6.13 TOST (1-α)100% CI for equivalence of testing modality means for $\alpha = 0.05$. Equivalence interval of $\pm 5\%$ of laboratory participant means. 129

6.14 Comparison of RMSE ($\mathcal{L}$) and participant age. 129

6.15 Objective MOS for each method in the evaluation set, averaged at each time-scale ratio. 131

7.1 [Colour Online] OMOQ system block diagram. Features coloured by group, detail shown for novel features. 137

7.2 Time-instance framing with reference anchor. Reference signal as top waveform and test signal ($\beta = 0.5$) as bottom waveform. Frames start at the same relative position in the signal. 139

7.3 [Colour Online] Phasiness features as functions of SMOS and TSM Ratio. Mean/Standard Deviation Phasiness No/Magnitude Weighting ([M/S]Ph[N/M]W). 143

7.4 [Colour Online] Spectral similarity features as functions of SMOS and TSM Ratio. 145

7.5 [Colour Online] Spectral similarity mean difference feature as functions of SMOS and TSM method. 145

7.6 [Colour Online] Spectral similarity mean difference feature as functions of TSM Ratio and TSM method. 146

7.7 [Colour Online] Transient features as functions of SMOS and TSM Ratio. 148

7.8 Neural network of proposed measure. Numbers denote number of layer output nodes, FC is a Fully Connected layer, LN is Layer Normalization, ReLU activation function, $\oplus$ is element-wise summation of layer input and output values and σ denotes a sigmoid activation layer. . . . 149

7.9 [Colour Online] Feature correlation matrix for final features. Absolute correlation for $\beta < 1$. 151

7.10 [Colour Online] Feature correlation matrix for final features. Absolute correlation for $\beta > 1$. 151

7.11 [Colour Online] Feature correlation matrix for final features. Average absolute correlation for $\beta < 1$ and $\beta > 1$ shown due to non-monotonic nature of relationship between features and time-scale ratio. 152

7.12 Box plot of best distance measure for each seed and training configuration ordered by median $\mathcal{D}$. PEAQB NN uses original PEAQ network, all others use the network described in Section 7.2.4. Lower is better, less spread means less reliance on initial seed. 153

7.13 Loss and Correlation for training, validation and test sets for each epoch. Best epoch shown as vertical line. 156

7.14 [Colour Online] Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined. 160

8.1 Proposed CNN dataflow. Kernel sizes in brackets, numbers denote layer size and number of channels, FC is a fully-connected layer, LN is layer normalisation, ReLU activation used unless specified. 170

8.2 Proposed RNN FF dataflow. D_F is feature depth, D_H is Hidden Dimensions, n is the number of directions, numbers denote layer sizes, FC is a fully-connected layer, LN is layer normalisation, ReLU activation used unless specified. 171

8.3 Proposed GRU FT network dataflow. D_F is feature depth, D_H is hidden dimensions, n is the number of directions, L is sequence length, numbers denote layer sizes, FC is a fully-connected layer and hashed sections are zero-padding to longest file in mini-batch. 172

8.4 Box plot of best distance measure for 30 seeds of each network configuration, ordered by median $\mathcal{D}$. $|X|$ is denoted by Mag, $\angle X$ by Ph, [MFCCs;D] by MD and hidden size denoted by 64 or 256. 173

8.5 [Colour Online] OMOS confusion matrix for CNN OMOQSE and OMOQDE. 176

8.6 [Colour Online] Training subset confusion matrix for CNN OMOQSE and SMOS. Test set shown as red dots. 177

8.7 [Colour Online] OMOS confusion matrix for BGRU-FT OMOQSE and OMOQDE. 178

8.8 [Colour Online] Training subset confusion matrix for BGRU-FT OMOQSE and SMOS. Test set shown as red dots. 179

8.9 Distribution of frames per MOS in training set. 179

8.10 [Colour Online] CNN estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined. 182

8.11 [Colour Online] Masked two-sample t-test for all CNN OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white. 183

8.12 [Colour Online] BGRU-FT estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined. 185

8.13 [Colour Online] Masked two-sample t-test for all BGRU-FT OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white. 186

List of Tables

3.1 Peak conditions for bin ranges for multi-resolution peak picking 45

4.1 MOS scores for subjective assessment of sound quality or impairment ITU-T (2019) 60

4.2 MOS scores seven grade comparison test ITU-T (2019) . . 60

5.1 Reference audio files with description. 87

6.1 Signal sources in each dataset class. All sources within a file are counted separately. 113

6.2 Mean MOS for Overall and Music, Solo Instrument, Voice classes of training source file. MOS for $\beta = 0.9961$ excluded. 123

7.1 RSME loss mean ($\overline{\mathcal{L}}$) and range ($\Delta\mathcal{L}$), PCC mean ($\overline{\rho}$) and range ($\Delta\rho$), median overall distance ($\widetilde{\mathcal{D}}$) and minimum overall distance ($\min(\mathcal{D})$). Trained to SMOS unless specified. Best results in bold. 154

7.2 Mean OMOS for each class of file and overall result. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$. 158

8.1 Test Loss ($\boldsymbol{\mathcal{L}_{te}}$) and PCC ($\boldsymbol{\rho_{te}}$), mean RSME loss ($\overline{\boldsymbol{\mathcal{L}}}$) and range ($\boldsymbol{\Delta\mathcal{L}}$), mean PCC ($\overline{\boldsymbol{\rho}}$) and range ($\boldsymbol{\Delta\rho}$), median overall distance ($\widetilde{\boldsymbol{\mathcal{D}}}$) and minimum overall distance ($\min(\boldsymbol{\mathcal{D}})$). Best single-ended results in bold. 175

8.2 Mean OMOS for each class of file and overall result for the proposed CNN OMOQ. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$. 181

8.3 Mean OMOS for each class of file and overall result for the proposed BGRU-FT network. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$. 184

Table of Abbreviations

Table 1: Abbreviations found within this Dissertation.

Abbreviation	**Description**
ADC	Analogue to Digital Converter
ADBB	Average Distorted Block (Basic)
AMS	Analysis Modification Synthesis
BGRU	Bidirectional Gated Recurrent Unit
BGRU-FT	Bidirectional Gated Recurrent Unit - Frame Target
BLSTM	Bidirectional Long Short Term Memory
CC - BY	Creative Commons By Attribution
CD	Compact Disc
CI	Confidence Interval
CNN	Convolutional Neural Network
DC	Direct Current (Zero Frequency)
DCT	Discrete Cosine Transform
DE	Double Ended
DFT	Discrete Fourier Transform
DI	Distortion Index
DIPL	Driedger's Identity Phase Locking Phase Vocoder

Continued on next page

Continued from previous page

Abbreviation	**Description**
DSP	Digital Signal Processing
EHSB	Harmonic Structure of Error (Basic)
ESOLA	Epoch Synchronous Overlap-Add
FC	Fully Connected
FCNN	Fully Connected Neural Network
FESOLA	Fuzzy Epoch Synchronous Overlap-Add
FF	Final Frame
FFT	Fast Fourier Transform
FT	Frame Target
FuzzyPV	Fuzzy Classification of Spectral Bins Phase Vocoder
GRU	Gated Recurrent Unit
GRU-FT	Gated Recurrent Unit - Frame Target
GUI	Graphical User Interface
HPSTrRat	Harmonic Percussive Separation Transient Ratio
HPTSM	Harmonic Percussive Separation Time-Scale Modification
IA	Instantaneous Amplitude
ICC	Intra-Class Correlation
IEEE	Institute of Electrical and Electronics Engineers
IP	Instantaneous Phase
IPL	Identity Phase Locking Phase Vocoder
IRCAM	Institut de Recherche et Coordination Acoustique/Musique
ITU	International Telecommunication Union

Continued on next page

Continued from previous page

Abbreviation	**Description**
ITU-R	International Telecommunication Union Recommendation
LLR	Log-Likelihood Ratio
LN	Layer Normalisation
LSEE-MSTFTM	Least Squares Error Estimation from Modified Short Time Fourier Transform
LSTM	Long Short Term Memory
MD	Mean Difference
MFCC	Mel-Frequency Cepstral Coefficient
MOS	Mean Opinion Score
MOV	Model Output Variable
MFPDB	Maximum Filtered Probability of Detection
MPhMW	Mean Phasiness Magnitude Spectrum Weighting
MPhNW	Mean Phasiness No Weighting
MS	Mid-Side
MSE	Mean Squared Error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
MedianOS	Median Opinion Score
NMFTSM	Non-Negative Matrix Factorization Time-Scale Modification
NMRB	Noise-to-Mask Ratio (Basic)
NN	Neural Network
ODG	Output Difference Grade
OLA	Overlap-Add

Continued on next page

Continued from previous page

Abbreviation	**Description**
OMOQ	Objective Measure of Quality
OMOQDE	Double Ended Objective Measure of Quality
OMOQSE	Single Ended Objective Measure of Quality
OMOS	Objective Mean Opinion Score
PCC	Pearson Correlation Coefficient
PEAQ	Perceptual Evaluation of Audio Quality
PEAQA	Perceptual Evaluation of Audio Quality Advanced
PEAQB	Perceptual Evaluation of Audio Quality Basic
PESQ	Perceptual Evaluation of Speech Quality
PSOLA	Pitch Synchronous Overlap and Add
PV	Phase Vocoder
PVSOLA	phase Vocoder with Synchronized Overlap-Add
PhaVoRIT	Phase Vocoder for Real-Time Interactive Time-Stretching
QSTI	Quasi-stationary Speech Transmission Index
RMS	Root Mean Square
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
SD	Sum and Difference
SE	Single Ended
SER	Signal to Error Ratio
SMOS	Subjective Mean Opinion Score
SNR	Signal to Noise Ratio
SOLA	Syncronous Overlap-Add

Continued on next page

Continued from previous page

Abbreviation	**Description**
SPC	Stereo Phase Correlation
MPhMW	Standard Deviation Phasiness Magnitude Spectrum Weighting
SPhNW	Standard Deviation Phasiness No Weighting
SPL	Scaled Phase Locking Phase Vocoder
SSMAD	Spectral Similarity Mean Absolute Difference
SSMD	Spectral Similarity Mean Difference
STFT	Short Time Fourier Transform
STI	Speech Transmission Index
STOI	Short-Time Objective Intelligibility
TOST	Two One Sided Test
TSM	Time-Scale Modification
TSMDB	Time-Scale Modification Database
uTVS	Mel-Scale Sub-band Modelling Time-Scale Modification
WAET	Web Audio Evaluation Toolkit
WSOLA	Waveform Syncronous Overlap-Add
WSS	Weighted Spectral Slope
ZFR	Zero Frequency Resonator

Table of Symbols

Table 2: Mathematics symbols found within this Dissertation.

Symbol	Description
$\overrightarrow{a}_c$	LPC for Original Speech
$\overrightarrow{a}_p$	LPC for Enhanced Speech
B	Stereo Balance
$B(f)$	Bark-Frequency
b	Bit depth
C	Stereo Phase Coherence
$\mathcal{D}$	Overall Model Distance
$\widetilde{\mathcal{D}}$	Median Overall Model Distance
D	MFCC Deltas
D'	MFCC Delta-Deltas
D_F	Feature dimension size
D_H	Hidden dimension size
d	Cohen's sample d
$E(\cdot)$	Expectation or Mean
F_o	Over-sampled sampling frequency

Continued on next page

Continued from previous page

Symbol	Description
F_s	Sampling frequency
f	Frequency (Hertz)
$G(q,k)$	Inter-rater reliability
k	Bin index
$\mathcal{L}$	Loss
$\boldsymbol{\mathcal{L}}$	Training, Validation, Testing Loss Vector
$\hat{\boldsymbol{\mathcal{L}}}$	Loss Distance
$\overline{\boldsymbol{\mathcal{L}}}$	Mean Loss
L	Sequence length
$M(f)$	Mel-Frequency
N	Frame length
n	Sample
$\boldsymbol{R}_c$	Auto-correlation Matrix of Original Speech
R_{xy}	Cross-Correlation of x and y
S	Frame shift
S_a	Analysis frame shift
S_s	Synthesis frame shift
T_s	Sampling period
t	Time
t_a^u	Analysis time instant for frame u
u	Frame number
$X(k)$	Fourier transform of x(n)
$\lvert X\rvert$	Magnitude of x(n)
$\lvert X\rvert^2$	Power of x(n)
$\angle X$	Phase of x(n)

Continued on next page

Continued from previous page

Symbol	**Description**
$x(n)$	Discrete signal
x_i	Opinion score of subject i
$\bar{x}_s$	Mean of all subjects in session s
$\bar{x}_{si}$	Mean score for subjective i in session s
Z_i	Normalised Subjective Score
α	Time-scaling parameter (Change in signal length)
β	Time-scale ratio (Change in playback speed)
γ	Scaled Phase Locking phase scaling factor
$\Delta\boldsymbol{\mathcal{L}}$	Loss Range
ΔP	Peak Delta
$\Delta\boldsymbol{\rho}$	PCC Range
ρ	Pearson Correlation Coefficient (PCC)
$\boldsymbol{\rho}$	Training, Validation, Testing PCC Vector
$\hat{\boldsymbol{\rho}}$	PCC Distance
$\overline{\boldsymbol{\rho}}$	Mean PCC
σ_s	Standard deviation for all subjects in session s
σ_{si}	Standard deviation for subject i in session s
$\sigma_{\Delta\varphi}$	Standard Deviation of Phasiness
$\Delta\varphi(u,k)$	Weighted Angle Difference
$\overline{\Delta\varphi}$	Mean Phasiness
$\widehat{\omega}_k(t_a^u)$	Instantaneous frequency
$[\cdot\ ;\ \cdot]$	Concatenation

Part I

Introduction

Chapter 1

Introduction

1.1 Background

Time-scale modification (TSM) of audio is a widely used process in applications ranging from data compression and reading for the blind, to post-production sound editing and musical composition. In each of these cases individual control over time and pitch can be beneficial. However, time and frequency are proportional; doubling the playback speed will double the frequencies contained in the signal. Digital Signal Processing (DSP) however, allows for the separation of time and frequency components, and gives the ability to scale a signal temporally, yet retain the frequency content of the recording. This is known as time-scaling with the inverse operation of manipulating the frequency domain, while retaining the signal duration is known as pitch-shifting.

In a general sense, TSM algorithms manipulate the temporal domain of a signal by framing the signal at a certain shift, and reconstructing the signal at a different shift. It is then the role of the TSM algorithm to maintain coherence within the signal. TSM algorithms can be

classified generally into frequency-domain, time-domain or source separation methods. Frequency-domain methods, modify magnitude and phase spectra, extracted using the Short-Time Fourier Transform to maintain coherence within the signal. Time-domain methods use different alignment calculations to ensure maximal similarity between the output signal and the current frame. Decomposition methods attempt to separate the signal into simpler components that are easier to time-scale, in an effort to increase the quality of TSM. Each approach has strengths and weaknesses, which will be discussed in detail within the dissertation.

Multi-channel TSM is an area with little published research. In general, novel TSM algorithms are presented for a single channel, with very few methods published that specifically consider multiple channels. The methods that have been published are frequency-domain methods and modify the phase spectrum to maintain phase relationships between channels. Due to the reliance on the phase spectrum, these methods are unable to be applied in cases where the phase spectrum is not used during time-scaling. Additionally, no qualitative or quantitative data has been published for these methods.

Subjective testing is the 'gold standard' when determining the quality of speech and audio. However, this testing is costly and time consuming. As a result, objective measures of quality have been developed that estimate subjective quality ratings. In some cases, these measures model human perception of sound, such as Perceptual Evaluation of Audio Quality (PEAQ) of Thiede et al. (2000) and Perceptual Evaluation of Speech Quality PESQ of Rix et al. (2001). These objective measures of quality decrease development time of algorithms and allow for greater comparison of quality between algorithms. However, measures of qual-

ity are generally developed for specific contexts, such as speech quality, speech intelligibility, audio codec development or blind source separation of audio. Currently there is no objective measure of quality suitable for determining the quality of time-scaled signals.

In this dissertation, we begin by presenting work improving stereo TSM and speech TSM. These improvements are then used in the development of multiple objective measures of quality. A dataset of time-scaled signals is collated and subjectively evaluated to form a quality ground truth that objective measures can be built on. The aim of these measures is to effectively estimate subjective quality ratings, thus providing a simple method for comparison of novel and previously published TSM algorithms. A doubled-ended measure applies hand-crafted features to a fully-connected neural network, while two single-ended measures use different deep-learning architectures to create data-driven features that are used as input. The performance of these measures is presented and the measures are used to evaluate the quality of TSM implementations.

1.2 Dissertation Organisation

This dissertation is organised into three parts. The first is the introduction. Following this, Part II is a literature review of signal processing fundamentals, TSM methods and measures of quality. Part III constitutes the research section of this work, and is concerned with improvements to TSM methods, the TSM dataset with subjective labels, the first objective measure of quality for time-scaled audio and two single-ended objective measures. Conclusions are drawn at the end of each research chapter, with overall conclusions and future research presented

in Part IV. Appendices can be found in Part V. A more detailed overview of each of these chapters is as follows.

- *Chapter 2* provides background in the fundamentals of signal processing required for understanding the work presented in this dissertation. Fundamentals of sound perception are covered, including perceptual frequency and loudness scales. Relevant elements of sound capture are covered, focusing on stereo microphone techniques and analogue to digital conversion. Finally, fundamental digital signal processing concepts are covered, including framing, windowing, the spectral domain and the Analysis Modification Synthesis framework.

- *Chapter 3* provides an in-depth exploration of TSM methods. Beginning with a general introduction, time-domain, frequency-domain and source separation methods are explored. Discussion is focused around methods used within this dissertation. A brief discussion of published methods of stereo TSM is also included.

- *Chapter 4* provides exploration and discussion of subjective and objective quality evaluation. Grading scales and normalisation of subjective results are discussed, followed by methods of subjective testing, and previous subjective testing of time-scaled signals. Various objective measures are discussed with a focus on Perceptual Evaluation of Audio Quality and Perceptual Evaluation of Speech Quality. Finally, figures of merit for objective measures are discussed.

- *Chapter 5* details improvements to TSM methods. The first section proposes two methods for improving the handling of stereo files within TSM. These offline and online methods are presented

along with objective and subjective evaluation of the proposed and previous art. The second section proposes an improvement to the Epoch Synchronous Overlap-Add TSM algorithm. Cross-correlation of time-smeared epochs is used to improve the quality of time-scaling for signals with modulating fundamental frequencies. Comparative and individual subjective results are presented.

- *Chapter 6* details the creation, subjective testing and evaluation of a dataset of time-scaled signals. The dataset, based on 108 reference files, covers a wide range of sound sources, time-scales and variety of TSM methods. Large scale subjective testing is used to evaluate the quality of the audio files, with statistical analysis applied to qualify the subjective testing. No specific mention is made of preference towards a particular TSM method, due to the nature of the testing, however general performance is compared to allow for future comparison of objective measures. The work of Chapter 5 is used here, with this chapter laying the groundwork for the remaining chapters.

- *Chapter 7* explores the development of the first effective objective measure for time-scaled audio. This measure extends the work of Thiede et al. (2000), through alignment of time-scaled signals. Novel features specific to TSM are introduced and significantly improve the performance of the objective measure. The hand-crafted features discuss within this chapter as used as input to a fully-connected neural network that is trained to the mean and median subjective scores of Chapter 6. Statistical analysis is performed on alignment methods, network architectures, objective quality estimates, and on estimates of quality for 16 different TSM implementations.

- *Chapter 8* continues the work of Chapter 7 by replacing the hand-crafted features with two deep learning-based approaches for features generation. A convolutional neural network and a bidirectional gated recurrent unit network are used independently to generate data-driven features that are fed to a fully-connected network for estimation of the subjective scores from Chapter 6. Comparison is made between quality estimates from the two single-ended objective measures, and subjective and objective scores from Chapter 7.

- *Chapter 9* provides a summary of all findings and conclusions within the dissertation, and concludes with directions for future research.

1.3 Contributions

All the work reported in this dissertation is based on papers listed in Section 1.4, with the specific research contributions being:

- A generalised method for time-scale modification of stereo files is proposed (Chapter 5).

- Methods for assessing the quality of stereo time-scale modification of stereo files are proposed (Chapter 5).

- Improved quality of speech time-scaling over previously published methods is proposed (Chapter 5).

- A dataset of varied sound sources including music, solo percussive and harmonic instruments, voice, and complex sound effects is presented (Chapter 6).

- The reference dataset is time-scaled by multiple methods and time-scales, forming a time-scaled dataset that allows for comparison with future time-scale modification methods (Chapter 6).
- The whole dataset is then rated subjectively and presented for use in future work (Chapter 6).
- The first effective objective measure for time-scale modification of audio is proposed (Chapter 7).
- The objective measure is evaluated against subjective opinion scores (Chapter 7).
- Time-scale modification methods are compared using the objective measure (Chapter 7).
- Single-ended objective measures that do not require a reference signal are proposed and evaluated (Chapter 8).
- Time-scale modification methods are compared using the single-ended objective measures (Chapter 8).

1.4 Publications

1.4.1 Publications Resulting from Dissertation Research

1. T. Roberts and K. K. Paliwal. Stereo time-scale modification using sum and difference transformation. In 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS), pages 1–5. IEEE, 2018.
2. T. Roberts and K. K. Paliwal. Time-scale modification using fuzzy epoch-synchronous overlap-add (FESOLA). In 2019 IEEE Work-

shop on Applications of Signal Processing to Audio and Acoustics, pages 31–34. IEEE, 2019.

3. T. Roberts and K. K. Paliwal. A time-scale modification dataset with subjective quality labels. The Journal of the Acoustical Society of America, 145(5), pages 3095–3103, 2020.

4. T. Roberts and K. K. Paliwal. An objective measure of quality for time-scale modification of audio. Under review with: The Journal of the Acoustical Society of America.

5. T. Roberts, A. Nicolson and K. K. Paliwal. Deep learning-based single-ended objective quality measures for time-scale modified audio. Under review with: The Journal of the Acoustical Society of America.

1.4.2 Additional Publications

1. T. Roberts and K. K. Paliwal. Frequency dependent time-scale modification. In 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS), pages 1–5. IEEE, 2018.

2. T. Roberts. A time-scale modification dataset with subjective quality labels, 2020. http://dx.doi.org/10.21227/ny9p-rv41. http://ieee-dataport.org/1987.

No assistance was received in the pursuit of the research and preparation of the dissertation beyond supervision, and suggestions of possible research directions and proof reading of Chapter 8.

1.5 Ethical Clearance for Experimentation

Throughout this work, the results of subjective experiments have been presented. These experiments involved volunteers listening to recordings and making a choice based on their evaluation of quality. All experiments were conducted with approval from the Griffith University Human Research Ethics Committee: database protocol number 2018/671.

Part II

Literature Review

Chapter 2

Fundamentals

2.1 Sound Perception

Human perception of sound is a complex task that relies on both physiological and psychological processes. The physiological process of transducing pressure waves to neural signals occurs in the ear, while the brain conducts the psychological processing of the neural signals. The ear can be sectioned into three main sections: the outer ear, middle ear and inner ear, shown in Figure 2.1.

The outer ear consists of the pinna and the ear canal. The purpose of the pinna is to amplify and direct sound waves into the middle ear. Wiener and Ross (1946) found that the pinna gives a gain of 3-5 dB at most frequencies, but peaks at 20dB for 1500Hz. The asymmetric shape of the pinna also assists in localisation of sounds (Ballou, 2008).

The middle ear begins with the eardrum which is an acoustic to mechanical transducer membrane. The main purpose of the middle ear is to match the impedance of the outer ear to the impedance of the

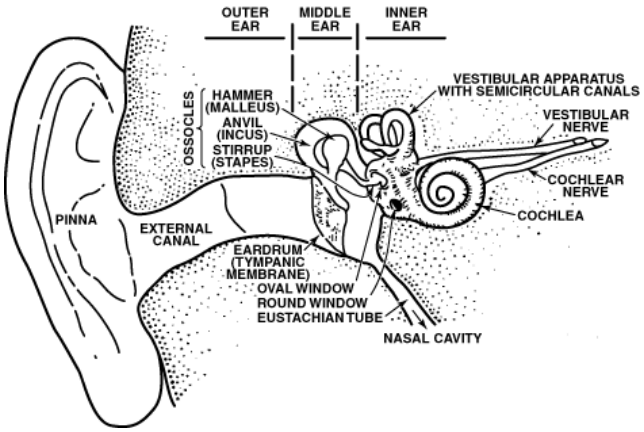

Figure 2.1: The external and internal ear. (Flanagan, 1972) *Reprinted with permission.*

inner ear. Sealed to the outside world, except while swallowing when the eustachian tube opens to balance the air pressure, the middle ear contains three bones known as ossicles. The ossicles, malleus, incus and stapes, transmit the acoustic energy from the ear drum to the inner ear through piston, lever and buckling motions (Pickles, 1988). Additionally, in the middle ear are two muscles, the tensor tympani and the stapedius muscle, which stiffen when a very loud sound is heard (more than 75dB above the hearing threshold) to protect the inner ear. This protection however is only effective for slow onset sounds with frequencies below 1.2kHz (Durrant and Lovrinic, 1995).

The inner ear contains two sections, the vestibular system (balance) and the auditory system (hearing), and is filled with fluid. The auditory system consists of the snail-shaped cochlea, which contains the basilar membrane that supports tiny hairs. These hairs are sensitive to different frequency ranges depending on their location, with high frequency hairs closest to the middle ear and low frequency hairs further away, and transmit electrical signals to the brain through the auditory nerve.

Due to the makeup of the ear, the frequency response is not consistent across the 20-20KHz audible range. The Fletcher-Munson equal-loudness curve, Figure 2.2, shows the perceived loudness for the audible range. For example, a 1kHz tone at 50dB SPL will be perceived to be the same loudness as an 80Hz tone at 75dB SPL. The shape of the curves indicates increased sensitivity to frequencies from 500 to 4000 Hz, which corresponds to the frequency range with the highest energy range for human speech.

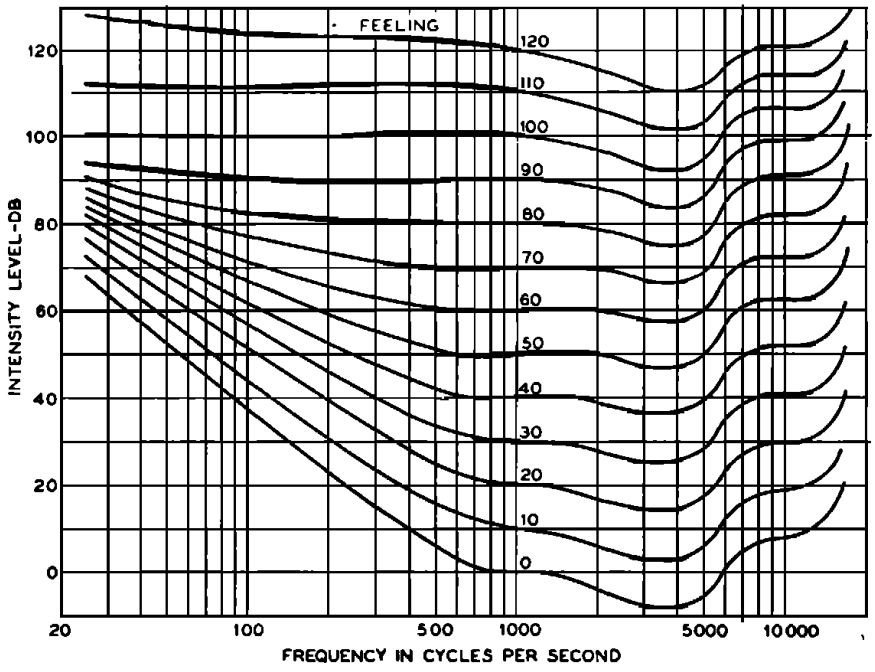

Figure 2.2: Fletcher-Munson equal-loudness curve. (Fletcher and Munson, 1933) *Reprinted with permission.*

Two units are used for the loudness of tones, phon and sone. Phons are the SPL in decibels of a tone, such that 40dB is 40 phons. Sones are designed to give a linear scale for loudness, rather than the logarithmic scale of phons (Stevens, 1936). An increase of 10 phons produces a

doubling in sones. Conversion between the two units uses

$$N = \left(10\frac{L_N - 40}{10}\right)^{0.30103} \tag{2.1}$$

from Fastl and Zwicker (2006), where L_N is the loudness in phons for $L_N > 40$phon. For frequencies other than 1kHz, the loudness in phons must be calibrated to the equal loudness curves before conversion to sones.

As with the non-linear nature of loudness perception, frequency perception is also non-linear. Stevens et al. Volkmann et al. (1937) first published the logarithmic nature of frequency perception and described the cochlea as a bank of overlapping, logarithmically spaced filter banks. The filter bandwidth increases with frequency as does the just noticeable difference between two tones. The Mel scale, proposed by Volkmann et al. (1937), and the Bark scale, proposed by Zwicker (1961) are the two commonly used scales based on these filter banks. Frequencies in Hertz (f) can be converted to Mel and Bark scales using

$$M(f) = 1125\ \log_e\left(1 + \frac{f}{700}\right) \tag{2.2}$$

and

$$B(f) = 6\log_e\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right). \tag{2.3}$$

The difference between these scales as a function of frequency in Hertz can be seen in Figure 2.3

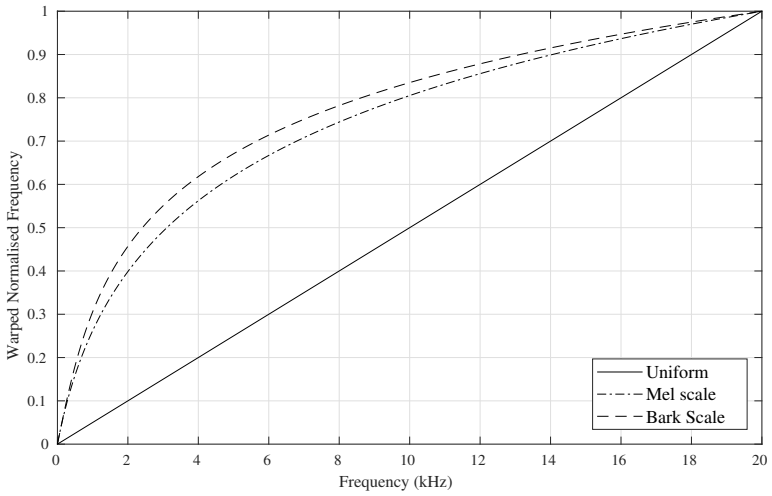

Figure 2.3: Comparison between Uniform Frequency Scale and Frequency Warped Mel and Bark Scales.

2.2 Sound Capture

The primary method of sound capture is through the use of microphones to convert acoustic pressure waves to electrical signals. The most basic configuration is that of a single microphone, which produces a single channel signal, also known as mono. Multiple stereo microphone techniques have been developed such as XY, ORTF and Mid/Side (MS). XY stereo configurations use a coincident pair of cardioid microphones oriented 90 degrees to each other. ORTF also uses two cardioid microphones, however the angle between microphones is 110 degrees with a microphone spacing of 170mm between the capsules. MS configurations also use two microphones, however one of the cardioid microphones is replaced with a figure-8 pickup pattern. The figure-8 is placed perpendicular and coincident to the cardioid. By duplicating the signal from the figure-8 microphone, and inverting the phase of one of the signals, the 'side' information can be captured. This allows for balancing of

the perceived stereo width of a recording by the balance of the cardioid (mid) and figure-8 (side) signals and offers advantages when considering stereo TSM processing. M/S encodes the stereo signal as sum and difference signals rather than left and right. This results in a mono compatible stereo signal with one channel containing the 'middle' information and the other containing the 'side' information. These signals are out of phase with each other (Ballou, 2008). This technique is also used in stereo FM broadcast (Ryan and Frater, 2002) and is discussed further in Section 3.6.1. Many more microphone techniques have been developed, but they are outside the scope of this review. Alternative capture methods have also been developed such as guitar pickups that use magnetic fields to convert the vibration of metal strings to electrical signals, piezo-electret microphones that use deformation in a crystal structure, and electronic instruments such as synthesizers that generate the electrical signals directly.

2.2.1 Analogue to Digital Conversion

Most signals of interest, are by nature analogue and continuous and, require conversion to a digital format if a computer is to be used to process the signal. This three-step procedure of sampling, quantizing and coding is known as analogue to digital conversion (ADC) and can be seen in Figure 2.4. By sampling a continuous-time signal $x_a(t)$, a discrete-time signal $x(n)$ is created at a series of discrete-time instants. The time between sampling is known as the sampling period (T_s), and is the inverse of sampling frequency, using $T_s = \frac{1}{F_s}$. The sampling frequency is also known as the sample rate. Quantization converts the continuous values of the sampled signal to discrete values $x_q(n)$. Due to the discrete nature of the quantized signal, each sample contains quantization error, which

is the difference in sample value before and after quantization. Coding is the final ADC process in which each quantized level is given a b-bit binary representation (Proakis and Manolakis, 2007).

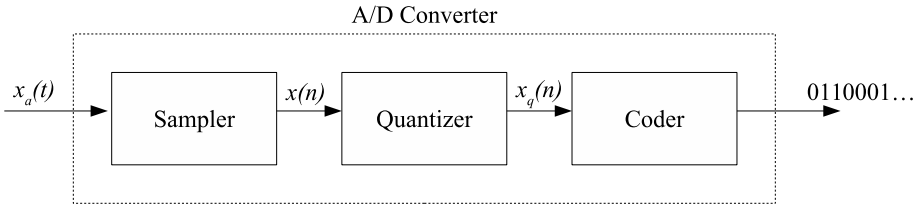

Figure 2.4: Analogue to Digital Converter Block Diagram.

In order for the analogue signal to be perfectly reconstructed without aliasing, the Nyquist-Shannon sampling theorem states that if a signal with a maximum frequency F_{max} is sampled at a sampling frequency of greater than $2F_{max}$, the original signal can be exactly recovered (Proakis and Manolakis, 2007). F_{max} is also known as the Nyquist rate.

Sampling rate varies based on the use case with speech often sampled at 8kHz or 16kHz, while CD quality audio has a sample rate of 44.1 kHz. Telephone systems band-limit speech to between 300 and 3400 Hz without significant impact on perceived quality (Huang et al., 2001). The 44.1 kHz sample rate of CD quality audio enables the full bandwidth of human hearing to be fully represented. Higher sample rates are used in some circumstances with sample rates of up to 192kHz supported on new audio hardware, however the extra processing cost and required down-sampling for distribution limits the usefulness of high sample rates.

2.3 Digital Signal Processing

2.3.1 Time Domain

Sound waves are represented in the time-domain, by the variation in pressure from the ambient pressure at a particular observation point as a function of time. Fig 2.5 shows the time-domain representation of male speech for the utterance "I am sitting in a room".

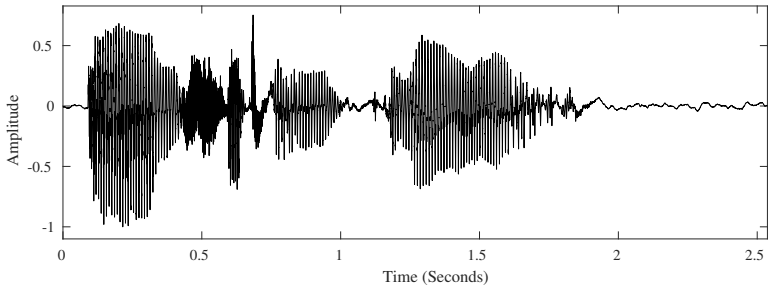

Figure 2.5: Waveform for a male utterance of "I am sitting in a room."

Framing

Audio signals are highly non-stationary processes; however the properties change very little over small durations of 20-40ms. As a result, it is possible to segment a signal into smaller sections, known as frames or windows, to create a quasi-stationary signal such that analysis can be undertaken. Frames are defined by two parameters, frame length and frame shift. Frame length (N) is the number of samples in each frame, while frame shift (S) is the number of samples between the beginning of each frame. Letting u denote the frame number, starting at 0, each frame will contain samples from uS to $uS + N - 1$. This process can be seen in Figure 2.6.

As the length of the frame used to analyse the signal increases,

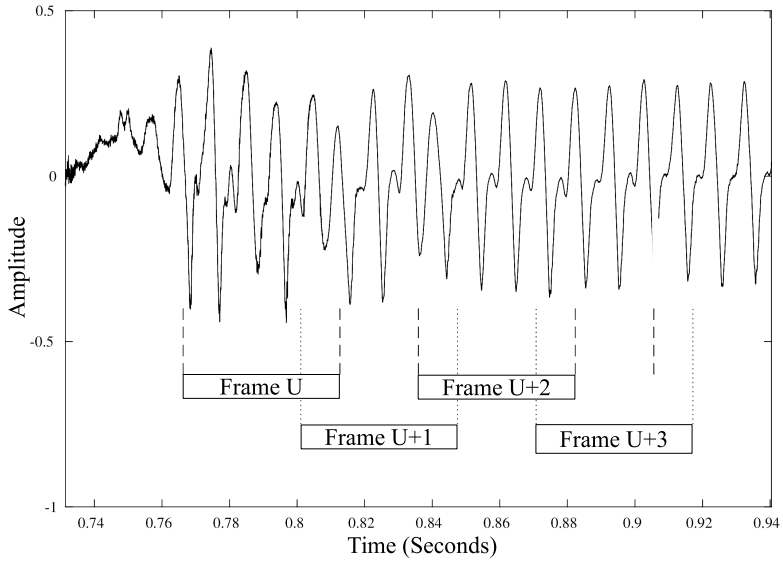

Figure 2.6: Framing an audio signal.

the frequency resolution increases, however the temporal resolution decreases. Conversely if the length of the frame is decreased, the temporal resolution increases, but the frequency resolution decreases. When determining optimal values for frequency-domain TSM, this is a major consideration as transients within the signal become smeared at longer frame lengths. Zero-padding of the signal is often used to increase the sampling of the frequency spectrum, but it does not increase the resolution of the frequency spectrum (Proakis and Manolakis, 2007).

Windowing

Windowing is the process of applying a function to a frame of data. The most basic window is a rectangle window, given by

$$w_{rect}(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

where N is the length of the frame and window function and n is the index. Use of this window results in discontinuities at the frame boundaries, which contributes to spectral leakage. This problem can be reduced by using tapered windows such as Hamming, Hann, Blackman, Kaiser and Chebyshev.

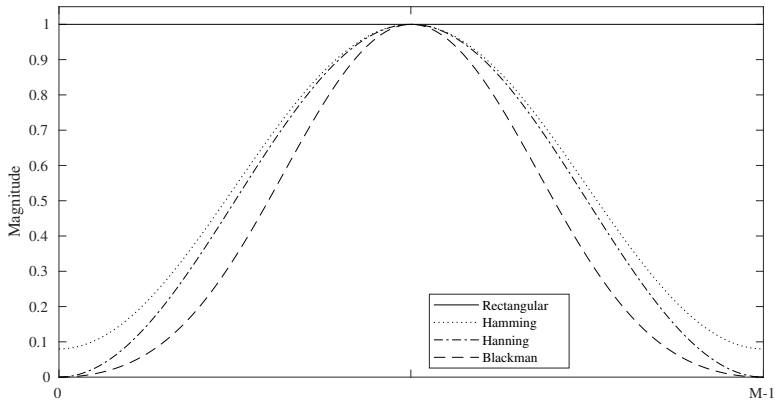

Figure 2.7: Shape of common window functions.

Research has been undertaken into the tonal effects of different windowing functions. Four different window functions were compared by Xiao and Jiang (2013) for their ability to maintain the quality of audio after pitch shifting using the phase vocoder. It was found that the Hann window had good performance in mid to high frequency bands, the Hamming window excelled in low frequencies and the Blackman window was suitable for all frequency ranges. The Hann window (often described as Hanning due to confusion with the Hamming window),

$$w_{hann}(n) = \frac{1}{2}\left(1 - cos\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{2.5}$$

allows the use of $\frac{N}{4}$ frame shifts for unity summation during overlap-add signal synthesis. If the frame shift was decreased to $\frac{5N}{6}$ the Blackman window could be used effectively.

2.3.2 Spectral Domain

The spectral domain is used to give a representation of the frequency and phase information within a signal. To do so, a frequency domain transform is applied to the time-domain signal. The Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) are the most common forms of frequency transform. The spectrum, X_ω, is often expressed in polar form,

$$X_\omega = |X_\omega| e^{j\angle X_\omega} \tag{2.6}$$

where $|X_\omega|$ is known as the magnitude spectrum and $\angle X_\omega$ is known as the phase spectrum.

The magnitude spectrum is a measure of the energy contained within each frequency band or bin of the spectrum, with the phase spectrum representing the normalised phase shift of each band or bin. Computation of the signal's instantaneous frequency within each bin is covered in Section 3.3.1.

Short Time Fourier Transform

The Fourier transform and subsequent Fast Fourier Transform (FFT) are fundamental tools in the field of digital signal processing, as they allow for transformation between the time and frequency domains. The discrete Fourier transform (DFT) is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}, \quad 0 \leq k < N-1 \tag{2.7}$$

and its inverse is given by

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\frac{j2\pi kn}{N}}, \quad 0 \leq n < N-1 \tag{2.8}$$

In practice the DFT is implemented as a Short-Time Fourier Transform (STFT) to make use of the quasi-stationary nature of speech and musical signals. The STFT applies an FFT to time-domain frames that have been modified with a window function. Portnoff (1976) showed that the STFT can be modified to the form of a DFT and implemented using the FFT algorithm. This modification requires the use of a window function such that $h(0) = 1$ and $h(n) = 0$ for $n = \pm 1N, \pm 2N, \pm 3N, etc.$ This ensures that $y(n) = x(n)$ for all values of n in the identity system (Portnoff, 1976). The computational complexity can also be reduced through a circular shift of half the length of the frame to avoid complex multiplications (Portnoff, 1976).

By framing the signal and computing the STFT it is possible to visualise the signal in the spectral domain using a spectrogram. Figure 2.8 shows the magnitude spectrum of the male utterance from Figure 2.5. Time is shown as the horizontal axis, frequency as the vertical axis and intensity shown by darkness.

2.3.3 Analysis Modification Synthesis Framework

The Analysis Modification Synthesis (AMS) framework provides a well-used method for signal processing in the frequency domain and is made of three sections, analysis, modification and synthesis, shown in Figure 2.9. In the initial section the time-domain signal is framed, windowed, the STFT is applied and the magnitude and phase spectra are calculated.

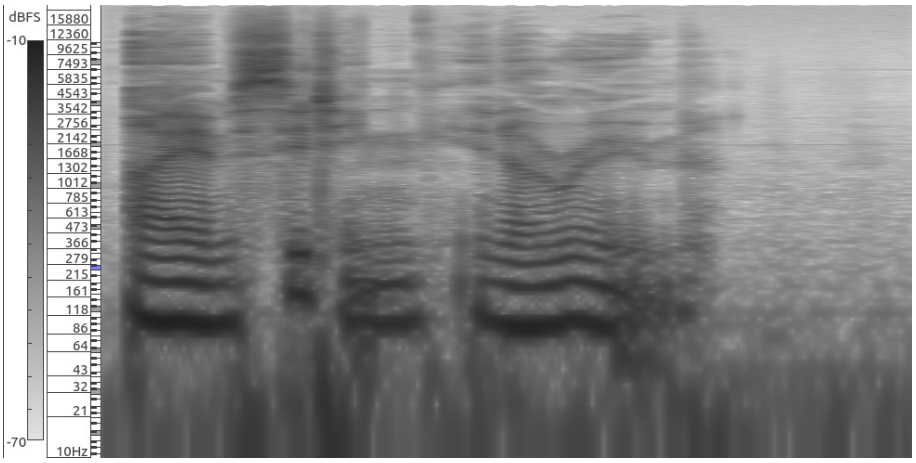

Figure 2.8: Spectrogram for male uttering "I am sitting in a room."

Modification is made to the appropriate spectra in the time-frequency processing. In TSM, time-scaling occurs in this section. Finally, an Inverse STFT is performed on the modified magnitude and phase spectra for conversion back to the time-domain. The final output signal is then reconstructed by summing overlapping frames by $N - S$, in a process known as overlap-adding.

Due to the overlapping windowed frames during the synthesis section, two methods can be used to normalise the resulting signal. One option is to normalise the signal using a signal constructed from windows. Alternatively, by using a window function such that the weighted signal sums to one, normalisation can be avoided. This is the preferred method for real-time implementation of TSM, however both options are useful in offline implementations.

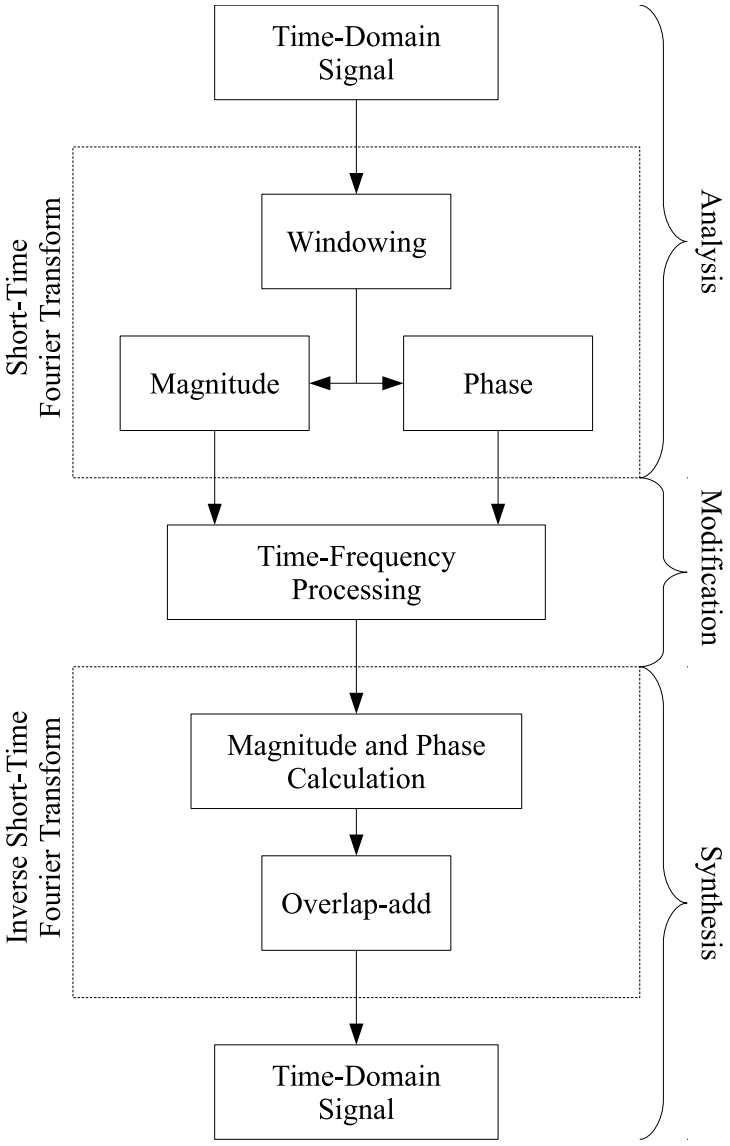

Figure 2.9: Analysis Modification Synthesis Algorithm

Chapter 3

Time-Scale Modification

Time-scale modification is the process of modifying the temporal domain of a signal, without modifying the frequency domain. It has found use in areas including music production, language learning and speech recognition systems. It is usually accomplished with an AMS framework by manipulating the ratio between the input and output frame shift sizes. The scaling factor is denoted by α for the change in signal duration (Roucos and Wilgus, 1985), while β is the playback speed (Sylvestre and Kabal, 1992) and is defined by

$$\alpha = \frac{1}{\beta} = \frac{S_s}{S_a} \tag{3.1}$$

where S_a is the analysis (input) frame shift, S_s is the synthesis (output) frame shift and β is the time-scale ratio. Values of α greater than one will expand the time-scale, while values less than one compress the time-scale. Within this dissertation, β is used in discussion of application of TSM, while α is used as a parameter within algorithms. A β of 0.5 denotes half speed while β of 2 denotes double speed.

TSM methods can roughly be described in three categories. Time-domain methods, frequency-domain methods and source separation methods. Time-domain methods operate directly on the signal, while frequency-domain methods makes use of filter banks and frequency transforms. Source separation methods decompose the signal into simpler signals, in order to use the strengths of time- and frequency-domain methods. Of note is the absence of time-scale modification incorporating machine learning. This is likely due to a lack of an objective measure of quality and a dataset of ideal time-scale modification. The concept of ideal TSM is explored in Section 6.1.

3.1 Historical Methods

Manipulating the temporal domain and pitch of an audio signals has a long history, beginning within music concrète. Initially these methods relied on varying tape speed to affect time and pitch, such that faster playback increases the pitch, while slower playback decreases the pitch. A digital version of this technique was also used within samplers, whereby signals are resampled before playback. It can be useful however to have independent control over time and pitch. For example, when fading between two signals of differing tempo, these methods may introduce unwanted pitch changes if played at the same tempo. An early analogue machine that allowed for time or pitch scaling of speech over a range of -40% to +10% was the Phonogène. It was a modified tape recorder that had multiple playback heads on a rotating drum. The absolute speed of the tape controlled the sound duration, while the relative speed to the rotating drum controlled the pitch (Zölzer et al., 2002).

3.2 Time Domain Methods

Time-domain methods are effective in time-scaling of speech, monophonic and percussive sound sources, and are computationally efficient algorithms. However, transient doubling and skipping are noticeable artefacts and phase jump artefacts give poor results when processing complex material. This is due to correlation processes preserving the most prominent periodic structures (Driedger and Muller, 2016). Time-domain methods are also known to produce an uneven speed due to tolerances required by the algorithm. Frame sizes are generally kept short, approximately 10ms, with tolerance values usually half the frame length (Driedger and Muller, 2016).

3.2.1 Overlap-Add

The most basic form of time-domain TSM is Overlap-Add (OLA). This method constructs the output by overlapping and adding windowed sections of the input waveform at a particular frame shift. Overlapping samples are summed, with the remaining samples concatenated to the end of the signal. This method, while being very simple to implement has multiple drawbacks. The biggest of these is that periodic structures in the waveform are not maintained. This results in distortion known as phase jump artefacts (Driedger and Muller, 2016). The artefacts are particularly damaging to signals with harmonic content where a warbling sound can be heard. Additionally, the pitch of the signal is not maintained. Conversely, percussive sounds are largely unaffected as they lack periodic structures. As a result, the uses for pure OLA are limited. However, OLA is used in the synthesis section of frequency-domain methods and extensions on this method are widely used.

3.2.2 Synchronised Overlap-Add

Synchronised Overlap-Add (SOLA), forms the basis for time-domain time-scaling techniques that preserve pitch. It was originally developed by Roucos and Wilgus (1985) for speech rate modification, with the aim of removing the iterative process required by the Least-Squares Error Estimation from the Modified Short Time Fourier Transform Magnitude (LSEE-MSTFTM) algorithm developed by Griffin and Lim (1984). SOLA time aligns successive windowed frames of the input waveform to the output before overlap adding. Time alignment is achieved by maximising the time-domain cross correlation between successive frames. Mathematically this can be represented by

$$y^o(n) = \frac{\sum_{u=-\infty}^{\infty} w^2(uS_s - n)x[n - u(S_s - S_a) - k(u)]}{\sum_{u=-\infty}^{\infty} w^2(uS_s - n)} \tag{3.2}$$

where w is the window function, u is the frame number and $k(u)$ is chosen to maximise the cross correlation.

This algorithm required moderate computation in 1985 and was capable of being implemented in a real-time system. Roucos and Wilgus (1985) included an algorithm for implementing their method and is as follows. The input signal $x_w(uS_s, n)$ is buffered into successive windows offset by S_a such that it takes the form of

$$x_w(uS_s, n) = w(uS_s - n)x[n - u(S_s - S_a)] \tag{3.3}$$

The system is then initialised according to

$$y(n) = w(n)x_w(0, n) \tag{3.4}$$

and

$$c(n) = w^2(n) \tag{3.5}$$

For $u = 1$ to the total number of frames, the value of $k(u)$ that maximises

$$R_{yx_w}(k) = \frac{\sum_{n=uS_s}^{uS_s+l} y(n)x_w(uS_s, n+k)}{[\sum_{n=uS_s}^{uS_s+l} y^2(n) \sum_{n=uS_s}^{uS_s+l} x_w^2(uS_s, n+k)]^{\frac{1}{2}}} \tag{3.6}$$

is calculated and denoted by k, where l is the length of the frame. The output is extended by incorporating the u^{th} window using

$$y(n) = y(n) + w(uS_s + k - n)x_w(uS_s, n+k) \tag{3.7}$$

and

$$c(n) = c(n) + w^2(uS_s + k - n) \tag{3.8}$$

Finally, the output waveform is normalised using

$$y(n) = \frac{y(n)}{c(n)} \tag{3.9}$$

for all values of n.

3.2.3 Waveform Similarity Overlap-Add

Waveform Similarity Overlap-Add (WSOLA) of Verhelst and Roelands (1993) extends the SOLA algorithm and the intermediate Time-Domain Pitch-Synchronised Overlap-Add algorithm (PSOLA) of Moulines and Charpentier (1990), by allowing for a time-scaling tolerance. WSOLA also extends these methods by calculating similarity with the natural progression of the input, rather than the output. The algorithm works in the following manner, with a graphical representation shown in Fig-

ure 3.1.

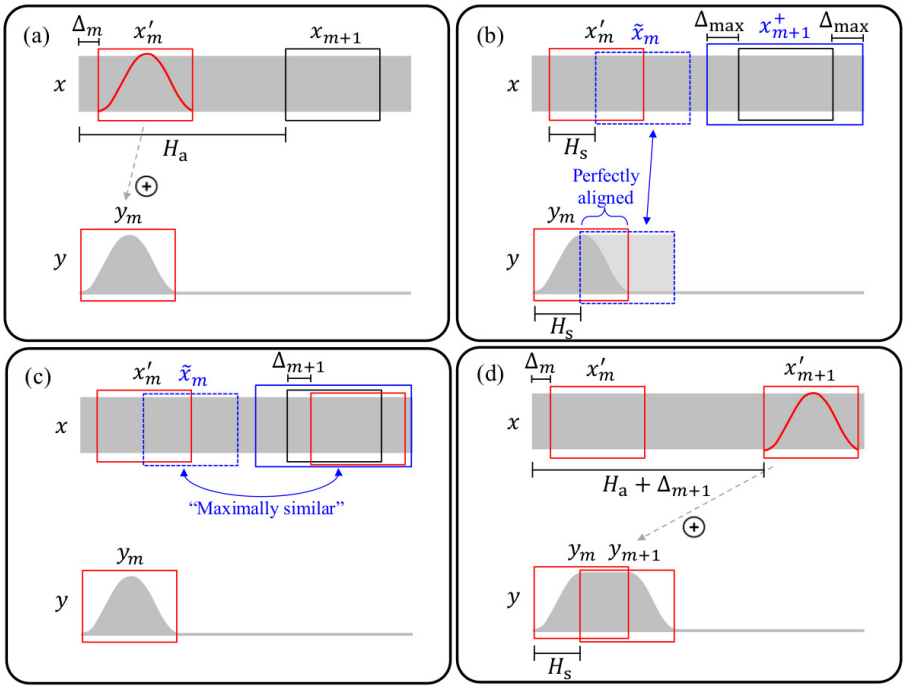

Figure 3.1: [Colour Online] WSOLA Algorithm. (Driedger and Muller, 2016) *Reprinted under CC BY.*

Within the diagram, H_a denotes the analysis frame shift, H_s is the synthesis frame shift and m is the frame number. The first frame x'_m is copied directly to the output and the next frame is determined based on the analysis shift size. This frame is extended by a tolerance factor $\Delta_{\max}$ to create x^+_{m+1} and the natural progression $\widetilde{x}_m$ from x'_m is determined. Cross correlation is then calculated between the natural progression and the extended frame. The frame position within the extended frame with the greatest cross correlation is then windowed and combined with the output using OLA. The process is repeated until the end of the file. As the output is constructed in a periodic manner it was found to be more algorithmically and computationally efficient than SOLA. The sound quality of this algorithm is also higher than that of SOLA.

3.2.4 Epoch Synchronous Overlap-Add

The production of voiced speech is a well understood process with air provided by the lungs passing through vocal chords, which close against each other in an oscillatory motion at the glottis (Huang et al., 2001). Significant excitation of the vocal system is generated at the moment of vocal chord closure (Rudresh et al., 2018). These moments are known as epochs or glottal closure instants. Figure 3.2 shows the location of these epochs within male speech calculated using the Zero Frequency Resonator (ZFR) algorithm. Note that the periodicity of the epochs matches the fundamental period of the signal. Generally, the related fundamental frequency or pitch of human speech lies between 60 and 300Hz (Huang et al., 2001).

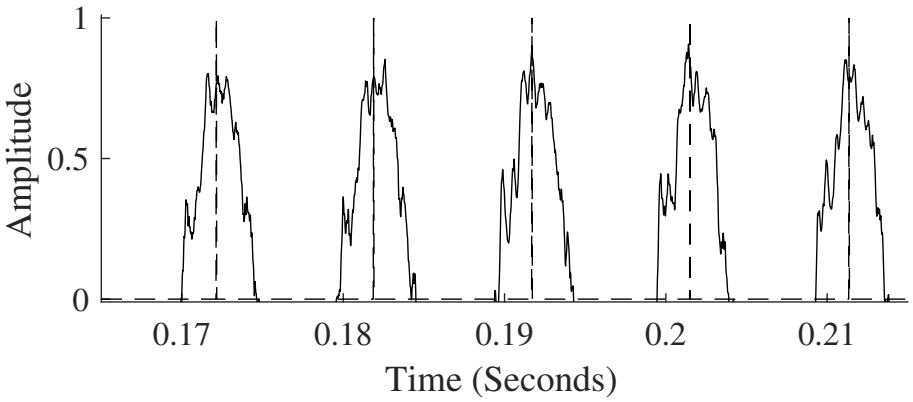

Figure 3.2: Epochs within male speech calculated using the Zero Frequency Resonator method.

Epochs can also be extracted from solo instrument recordings. Figure 3.3 shows epochs extracted from a recording of a solo flute. Note that the epochs are still aligned with the fundamental frequency of the waveform.

Epoch Synchronous Overlap-Add (ESOLA) of Rudresh et al. (2018) is motivated by the relatively small changes in fundamental frequency

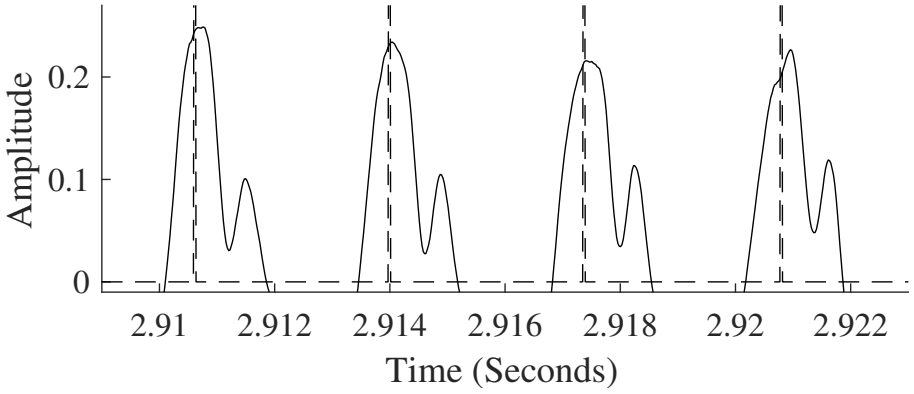

Figure 3.3: Epochs within a solo flute recording calculated using the Zero Frequency Resonator method.

across a range of speaking rates. As the fundamental frequency depends on the glottal closure instants, epochs make a logical candidate for re-aligning segments of the source file at a new time-scale. Multiple methods of producing epochs were considered, with the ZFR method used as it gives reliable estimates with a lower computational complexity (Murty and Yegnanarayana, 2008). The ZFR method is applied in the following manner. The signal is first pre-processed by calculating the first difference,

$$\hat{x}(n) = x(n) - x(n-1) \tag{3.10}$$

to remove any low frequency bias present in the signal.

$\hat{x}(n)$ is then passed through two ideal Zero Frequency Resonators,

$$y_1(n) = -\sum_{k=1}^{2} a(k) y_1(n-k) + \hat{x}(n) \tag{3.11}$$

and

$$y_2(n) = -\sum_{k=1}^{2} a(k) y_2(n-k) + y_1(n) \tag{3.12}$$

where $a(k)$ are filter coefficients.

The resulting trend in the filtered signal, $y_2(n)$ is removed through successive mean-subtraction operations,

$$y(n) = y_2(n) - \frac{1}{2N_{ZFR}+1} \sum_{m=-N_{ZFR}}^{N_{ZFR}} y_2(n+m) \tag{3.13}$$

where $2N_{ZFR}+1$ is chosen to be 1 to 2 times the fundamental pitch period. Finally, zero crossings in $y(n)$ indicate epochs within the signal.

Time-scaling of the source signal begins with pitch-blind windowing, with $\frac{N}{2}$ overlap. The synthesis overlap between frames is then increased or decreased for proportional changes in speed. To align frames, an output epoch frame is extracted starting at $S_s = \alpha u S_a$ and the location of the first epoch within the frame is determined. Similarly, an analysis epoch frame is extracted and next epoch after the location of the first output epoch is found. The difference in samples between epochs is then used as an offset when extracting the final synthesis sample and epoch frames. Frames are windowed using cross-fading functions before overlap-adding. An output signal of epochs is also maintained for alignment of future frames. ESOLA is very computationally efficient, and if the source signal is to scaled to multiple time-scale ratios, the epochs may be reused.

3.3 Frequency Domain Methods

Frequency-domain methods are preferred for harmonic sources and generate a higher quality output than time-domain methods for complex signals (Driedger and Muller, 2016), making use of the frequency domain and sum-of-sinusoids decomposition. However, these methods can produce poor quality output when processing transient source materi-

als. Many improvements have been developed to reduce the transient smearing and 'phasiness' artefacts of the Phase Vocoder. Frequency-domain methods can also provide constant ratio time-scaling, with no requirement for a tolerance factor.

3.3.1 Phase Vocoder

The Phase Vocoder (PV) was originally formulated to reduce transmission bandwidth, but the algorithm also allowed for time-scaling and pitch-shifting. These two applications are met by first representing speech signals by their short-time amplitude and phase spectra. Flanagan and Golden (1966) proposed using a bank of k band-pass filters where the output from each filter is the simultaneous amplitude and phase modulation of a carrier by the short-time amplitude and phase spectra (Flanagan and Golden, 1966) of the original signal, evaluated at the band-pass filter centre frequency. The phase spectra of this representation is not well bounded, however the derivative of the phase component is. The k^{th} filter output can be represented as

$$\widetilde{f}_k(t) = |F(\omega_k, t)| cos[\omega_k, t + \widetilde{\varphi}(\omega_k, t)], \quad \text{where} \quad \widetilde{\varphi}(\omega_k, t) = \int_0^t \dot{\varphi}(\omega_k, \tau) d\tau \tag{3.14}$$

The reconstruction of the signal consists of the sum of k oscillators modulated in phase and amplitude. Time compression and expansion is achieved through scaling the phase-derivative spectrum and scaling the playback speed by the same factor. Due to a lack of a requirement that the scaling factor be an integer, this method allowed for non-uniform changes in the time-scale (Flanagan and Golden, 1966). This method of using bandpass filters was superseded with the introduction of the digital Phase Vocoder of Portnoff (1976). The digital Phase Vocoder makes use

of the STFT rather than a bank of band-pass filters and results in more efficient calculation of amplitude and phase spectra. This method uses the AMS framework, described in Section 2.3.3, and is an identity system if no time scaling is applied. Equations from this point forward will be as per Laroche and Dolson (1999) for consistency throughout the Phase Vocoder and its improvements.

During the analysis phase, the input signal is first windowed with an overlap of at least 75% for Hann or Hamming windows, before being circular shifted by $\frac{N}{2}$, and transformed to the frequency domain. The resulting non-heterodyned STFT representation of the signal at analysis time instant t_a^u is given by

$$X(t_a^u, \Omega_k) = \sum_{n=0}^{N-1} h(n)x(t_a^u + n)e^{-j\Omega_k n} \tag{3.15}$$

Time instants are set according to the analysis shift size such that $t_a^u = uS_a$. The magnitude and phase of the STFT can then be calculated using

$$|X(t_a^u, \Omega_k)| = \sqrt{[\Re(X(t_a^u, \Omega_k))]^2 + [\Im(X(t_a^u, \Omega_k))]^2} \tag{3.16}$$

and

$$\angle X(t_a^u, \Omega_k) = tan^{-1}\frac{\Im(X(t_a^u, \Omega_k))}{\Re(X(t_a^u, \Omega_k))} \tag{3.17}$$

where $\Omega_k = \frac{2\pi k}{N}$ is the centre frequency of the k^{th} STFT bin.

During time-frequency processing $X(t_a^u, \Omega_k)$ is modified in two ways. The first is that the analysis shift size is different to the synthesis shift size, and the second is that the phase is modified such that phase coherence is maintained. Modification of the phase uses a technique known as phase unwrapping, where the instantaneous frequency of a sinusoid

is calculated from the phase increment between two consecutive STFT frames (Laroche and Dolson, 1999). The process is called phase unwrapping because the unwrapped synthesis phase is calculated from the wrapped instantaneous phase (Laroche and Dolson, 1999). The instantaneous frequency $\widehat{\omega}_k(t_a^u)$ is calculated by first calculating the heterodyned phase increment using

$$\Delta\Phi_k^u = \angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) - S_a\Omega_k \tag{3.18}$$

The phase increment Φ_k^u is reduced to between $\pm\pi$ by taking its principle determinate,

$$\Delta_p\Phi_k^u = \Delta\Phi_k^u - 2\pi(round(\frac{\Delta\Phi_k^u}{2\pi})) \tag{3.19}$$

Finally, the instantaneous frequency of the closest sinusoid to the centre frequency of the bin is calculated using

$$\widehat{\omega}_k(t_a^u) = \Omega_k + \frac{1}{S_a}\Delta_p\Phi_k^u \tag{3.20}$$

After calculating the instantaneous frequency, the synthesis phase can be calculated using the phase propagation equation:

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^{u-1}, \Omega_k) + S_s\widehat{\omega}_k(t_a^u) \tag{3.21}$$

Conceptually, the modification process calculates the phase progression for the analysis shift duration, calculates how far it should have progressed for the synthesis shift duration and adjusting accordingly. The modified frame is reconstructed according to

$$Y(t_s^u, \Omega_k) = |X(t_a^u, \Omega_k)|e^{\angle Y(t_s^u, \Omega_k)} \tag{3.22}$$

by using the original magnitude and the synthesised phase.

After time-frequency processing, the magnitude and phase spectra are transformed back to the time-domain, circular shifted by half, and an optional window function is applied, before overlap-adding to the output. The current input and output phases are stored for use with the following frame.

The Phase Vocoder offers several benefits over time-domain methods. Pitch shifting is easily achieved by scaling of the phase spectrum, there are improvements in the quality of time-scaling complex material and the time-scale of the processed signal is constant. However, the Phase Vocoder suffers from phasing when using non-integer α values, reverberation, loss of presence, and transient smearing artefacts as well as a higher computational cost.

3.3.2 Phase-Locking Phase Vocoders

A large amount research has been published on improving the quality of the Phase Vocoder, with a focus on reducing 'phasiness'. One such improvement is known as phase-locking. To understand why phase locking works, the Phase Vocoder algorithm must be considered in more detail.

The PV algorithm ensures that phase coherence is maintained within individual bins (horizontal phase coherence), but does not maintain phase coherence across multiple bins. This phase coherence between multiple bins is known as vertical phase coherence. For the resynthesised STFT to be valid, both the horizontal and the vertical phase coherence must be maintained (Laroche and Dolson, 1999). A loss of vertical phase coherence is heard as 'phasiness' or reverberation. Laroche and Dolson (1999) explore vertical phase coherence by first showing the relationship between the input and output phases depends on the ini-

tial analysis and synthesis phases, the current analysis phase, and the cumulative effect of any phase unwrapping errors. This dependence is shown in

$$\angle(t_s^u, \Omega_k) = \phi_s(0,k) + \alpha[\angle X(t_a^u, \Omega_k) - \angle X(0, \Omega_k)] + \alpha \sum_{i=1}^{u} 2\pi m_k^i \quad (3.23)$$

where $2\pi m_k^i = \Delta_p \Phi_k^u - \Delta\Phi_k^u$ is the difference between the phase increment and the principle determinate of the phase increment.

There are three conclusions that can be drawn from this. The first is that if an analysis phase is incorrectly estimated, this will not cause further phase drift in any subsequent frames provided that m_k^i is correct. The second is that any errors in phase unwrapping are cumulative. The third conclusion is that phase unwrapping errors that are multiples of $2\alpha\pi$ will not be heard. This is the reason that scaling with integer α values is of a higher quality than non-integer values of α. If α values other than integers are considered the phase coherence between nearby channels is only maintained if the "sums of the unwrapping factors $\sum_{i=1}^{u} 2\pi m_k^i$ are equal (modulo 2π) in nearby channels" (Laroche and Dolson, 1999). In a case such as noise, the phases for each channel will be random and phase coherence will be quickly lost. For signals other than noise, differences in phase will cause the vertical phase coherence to be gradually lost.

Phase locking was first proposed by Puckette (1995), and while the method produced improvement for simple test signals and improved vertical phase coherence, "informal listening tests show that the reduction in 'phasiness' is very signal-dependant and, unfortunately, never dramatic. Laroche and Dolson (1999), after exploring the reasons for the 'phasiness' as described above, proposed two methods of phase lock-

ing known as Identity Phase Locking (IPL) and Scaled Phase Locking (SPL). Both of these methods update the phases of the peak channels in the frequency spectrum, and lock the surrounding channel's phase propagation to the nearest peak. A channel is determined to be a peak if its magnitude is larger than its four nearest neighbours (Laroche and Dolson, 1999). The region to be locked to the peak can be determined in a few ways, but is most often set as the mid-point of two consecutive peaks. An alternative method is to use the local minima between peaks as the region boundary.

IPL finds peaks in the magnitude spectrum and updates the phases of only these channels according to the traditional Phase Vocoder technique described above. The rotation (θ) required for each channel in the peak's region of influence is then calculated according to

$$\theta = \angle Y(t_s^u, \Omega_l) - \angle X(t_a^u, \Omega_l) \tag{3.24}$$

where Ω_l is the centre frequency of the peak. This rotation for all phases can then be applied to the original frame by complex multiplication

$$Y(t_s^u, \Omega_k) = ZX(t_a^u, \Omega_k), \quad \text{where } Z = e^{j\theta} \tag{3.25}$$

As this method only requires the unwrapping of peak bins and one complex multiplication for each channel, there is a large increase in efficiency over the traditional Phase Vocoder.

SPL improves upon IPL by accounting for peaks moving between channels. If a peak moves from channel k_0 at time instant $u-1$ to channel k_1 at time instant u, equations 3.18-3.22 are updated to become

$$\Delta\Phi_k^u = \angle X(T_a^u, \Omega_{k_1}) - \angle X(t_a^{u-1}, \Omega_{k_0}) - S_a\Omega_{k_1} \tag{3.26}$$

$$\widehat{\omega}_{k_1}(t_a^u) = \Omega_{k_1} + \frac{1}{S_a}\Delta_p\Phi_k^u \tag{3.27}$$

$$\angle Y(t_s^u, \Omega_{k_1}) = \angle(t_s^{u-1}, \Omega_{k_0}) + S_s\widehat{\omega}_{k_1}(t_a^u) \tag{3.28}$$

and

$$\angle Y(t_s^u, \Omega_k) = \angle Y(t_s^u, \Omega_{k_1}) + \gamma[\angle X(t_a^u, \Omega_k) - \angle X(t_a^u, \Omega_{k_1})] \tag{3.29}$$

The previous peak location is determined by finding the region in the previous frame that the current peak belongs to. The peak of this region is then set as the relative previous peak. Equation 3.29 is the generalisation of the identity phase locking phase propagation equation where γ is a phase scaling factor. If γ is set to one, the equation is the equal to that of the identity phase locking. γ is normally set to a value of between one and α. There is little to support this (Laroche and Dolson, 1999), but it can be shown that with integer values of α and correct initialisation, phase differences are also scaled by $\gamma = \alpha$. Laroche and Dolson (1999) suggest that setting $\gamma \approx 2/3 + \alpha/3$ further helps to reduce phasing in the signal. In order to avoid $2\gamma\pi$ channel jumps in the synthesis phase, the analysis phases must be in an unwrapped state before applying Equation 3.29. This method requires more calculation than IPL, but gives a consistently higher quality of output (Laroche and Dolson, 1999).

3.3.3 Multi-Resolution Peak Picking and Sinusoidal Trajectory Heuristics

The Phase Vocoder for Real-Time Interactive Time Stretching (PhaVoRIT) of Karrer et al. (2006) makes three improvements to the Phase Vocoder algorithms of Laroche and Dolson (1999). The first is a change in the requirement for a channel to be labelled as a peak. As discussed

in Section 2.1, the human frequency perception is a non-linear system and the PhaVoRIT adjusts the finding of peaks in response to this. Instead of using a linear scale for determining peaks, multi-resolution peak-peaking is used. A channel is considered to be a peak based the increasing scale shown in Table 3.1 for a frame length of 4096 and a sample rate of 44.1kHz.

Table 3.1: Peak conditions for bin ranges for multi-resolution peak picking

Sub-band	**Bin Index**	**Freq. Range**	**Peak Condition**
1	0 - 16	$\approx$ 0 - 172 Hz	Always a peak
2	17 - 32	$\approx$ 183 - 345 Hz	$>$ 2 nearest neighbours
3	33 - 64	$\approx$ 355 - 689 Hz	$>$ 4 nearest neighbours
4	65 - 128	$\approx$ 700 - 1378 Hz	$>$ 8 nearest neighbours
5	129 - 256	$\approx$ 1389 - 2756 Hz	$>$ 16 nearest neighbours
6	257 - 512	$\approx$ 2767 - 5513 Hz	$>$ 32 nearest neighbours
7	$\geq$ 513	$>$ 5523 Hz	$>$ 64 nearest neighbours

This change increases the bass response of the time-scaled signal and reduces musical overtones (Karrer et al., 2006). The second change is a modification to the SPL. The PhaVoRIT algorithm takes into account that peaks and notes are not infinite in duration, with a method called sinusoidal trajectory heuristics. This technique improves SPL by stopping peak trajectories jumping across multiple bins. Figure 3.4 shows the trajectory of each peak for two consecutive frames. As can be seen there are cases when two peaks at the current time instant t_a^u, use the same peak from the previous time instant t_a^{u-1}. The synthesis phase for both of these new peaks is therefore linked to the same single previous peak. While this is less of an issue when the distance between the current and previous peak is small, it causes audible artefacts if the jump is large. The onset of a new note is a good example of when the phase

of a previous peak is unrelated to the new peak and should not be used to calculate the synthesis phase. Karrer et al. (2006) propose a system where the distance between a current and previous peak is used to determine if the previous peak belongs to the current peak. The entire spectrum is divided into side bands shown in Table 3.1. The allowed region for the sinusoid to jump is within its sub-band. For example, if the current peak is located at bin 86, bins 65 to 128 will be searched for the closest peak. If there is no peak within this range, the peak at bin 86 is considered to be a new sinusoidal trajectory and the synthesis phase is calculated using the standard phase propagation equations. This modification reduces blurred note onsets and high frequency warbling (Karrer et al., 2006).

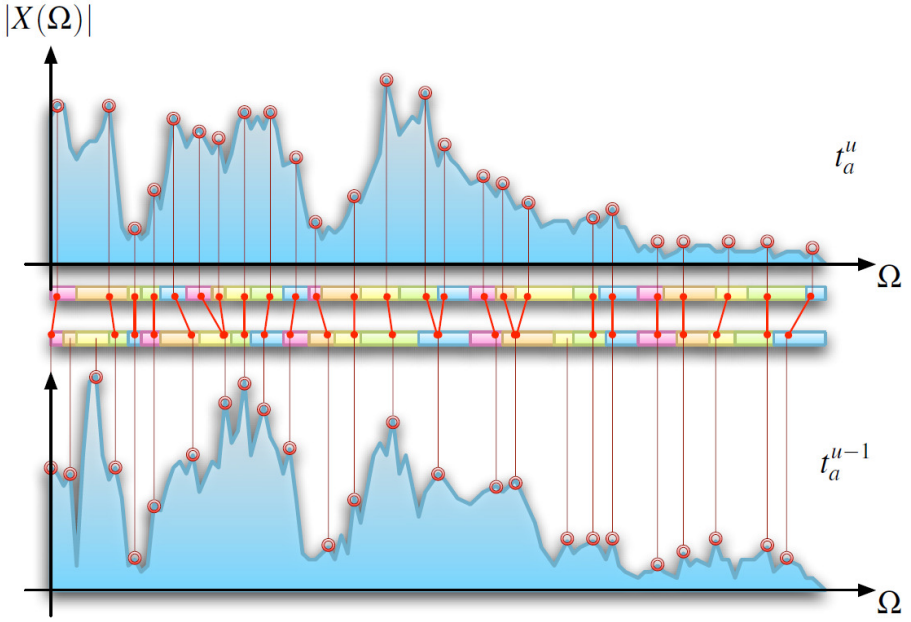

Figure 3.4: [Colour Online] Sinusoidal Trajectories from Karrer et al. (2006). *Reprinted with permission.*

Silent Passage Phase Reset

PhaVoRIT also implements silent passage phase reset, and is used to remove the cumulative long term errors due to incorrect phase unwrapping as discussed in Section 3.3.2 above. When the level of the signal drops below -21dB, it is assumed that the signal contains no meaningful information. When the signal energy rises back above -19dB, the synthesis phases are reset to their analysis phases. This results in a soft audible click, but is somewhat masked by the increase in signal level (Karrer et al., 2006). This process reduces reverberation in the output signal.

3.3.4 Fuzzy Classification of Spectral Bins

Time Stretching using Fuzzy Classification of Spectral Bins (FuzzyPV) of (Damskägg and Välimäki, 2017), is an extension of the IPL, and challenges the sum-of-sinusoids decomposition inherent in the STFT. This method considers each bin of the STFT to be a superposition of sinusoidal, transient and noise components. Consequently, spectral bins are given degrees of membership to tonalness, noisiness and transientness classes, resulting in a fuzzy classification for each bin.

As tonal components of a signal appear as horizontal ridges within a spectrogram, while transient components appear as vertical ridges, median filtering is used to extract tonal and transient spectra $X_s[u,k]$ and $X_t[u,k]$, using

$$X_s[u,k] = \text{median}(|X_s[u - \frac{L_t}{2} + 1, k]|, \ldots, |X_s[u + \frac{L_t}{2} + 1, k]|) \quad (3.30)$$

and

$$X_t[u,k] = \text{median}(|X_s[u - \frac{L_f}{2} + 1, k]|, \ldots, |X_s[u + \frac{L_f}{2} + 1, k]|) \quad (3.31)$$

L_t and L_f are the median filter lengths in time and frequency directions.

The tonal and transient STFTs are then used to estimate tonalness, transientness and noisiness using

$$R_s[u,k] = \frac{X_s[u,k]}{X_s[u,k] + X_t[u,k]} \quad (3.32)$$

and

$$R_t[u,k] = 1 - R_s[u,k] = \frac{X_t[u,k]}{X_t[u,k] + X_s[u,k]} \quad (3.33)$$

where transientness is the compliment of tonalness.

Finally, noisiness is calculated according to

$$R_n[u,k] = 1 - |R_s[u,k] - R_t[u,k]| = \begin{cases} 2R_s[u,k], & \text{if } R_s[u,k] \leq 0.5 \\ 2(1 - R_s[u,k]), & \text{otherwise} \end{cases} \quad (3.34)$$

as noise bins were found to be normally distributed around $R_s = 0.5$.

The IPL method of Laroche and Dolson (1999) is used to scale all bins. To avoid the metallic artefact from phase locking non-sinusoidal bins, psuedo-random noise ($d[u,k]$) is added to synthesis phase values based on the estimated noisiness of the bin, using

$$\angle Y'[u,k] = \angle Y[u,k] + \pi A_n[u,k](d[u,k] - \frac{1}{2}) \quad (3.35)$$

$A_n[u,k]$ is the phase randomisation factor and is calculated according to

$$A_n[u,k] = \frac{1}{4}[tanh(b_n(R_n[u,k] - 1)) + 1][tanh(b_\alpha(\alpha - \frac{3}{2})) + 1] \quad (3.36)$$

where b_n and b_α control the shape of the non-linear mapping. $b_n b_\alpha = 4$ was used in Damskägg and Välimäki (2017). $d[u,k]$ is drawn from a uniform distribution of the interval [0,1].

Transient preservation is similar to that of Röbel (2003), however it uses the previously estimated transientness. Transient detection begins with calculation of an overall transientness for each frame using

$$r_t[m] = \frac{1}{N-1} \sum_{k=1}^{N-1} R_t[u,k] \tag{3.37}$$

before the first backward difference approximates the derivative of the signal,

$$\frac{d}{dm} r_t[m] \approx \frac{1}{H_a}(r_t[m] - r_t[m-1]) \tag{3.38}$$

Transients are determined as local maxima above a given threshold.

To reduce transient smearing, magnitude spectrum bins with $R_t[u,k] > 0.5$ are suppressed, using

$$|Y[u,k]| = (1 - R_t[u,k])|X[u,k]| \tag{3.39}$$

Phase reset is then applied to those bins as the detected transient is centred within the analysis window, by using the analysis phase as the synthesis phase. Finally, the magnitudes of the transient bins are compensated to account for the loss of transient energy due to the transient appearing in adjacent frames.

Subjective evaluation showed showed similar average performance to Harmonic Percussive Separation Time-Scale Modification and Elastique. Quality was shown to be high for complex signals with easily discernible transients, however quality was poor for the Drum Solo signal tested.

3.3.5 Mel-Scale Sub-band Modelling

Multi-component Time-Varying Sinusoidal decomposition (uTVS) of (Sharma et al., 2017) bypasses the error prone phase unwrapping and quasi-stationary assumption of traditional frequency-domain methods, through direct calculation of Instantaneous Amplitude (IA) and Phase (IP). uTVS employs an AMS framework. Analysis begins by oversampling the input signal to $F_o = 6F_s$ and passing the resulting signal through a filter bank containing 32 filters, each with a tap length of 2048. Hann windowed sinc filters are used, with centre frequencies spaced linearly in the mel-scale. Modification is accomplished by taking each narrow band output from the filter bank and analytically calculating its IA and IP. First, let $x_{a,k}[n]$ be the analytic signal for $x_k[n]$, where k is the filter index. The Hilbert transform ($\mathcal{H}(\cdot)$) is then used in calculation of the analytic signal,

$$x_{a,k}[n] = x_k[n] + j\mathcal{H}(x_k[n]) \tag{3.40}$$

such that $a_k[n]$ and $\phi_k[n]$ can be estimated using

$$a_k[n] = |x_{a,k}[n]| \tag{3.41}$$

and

$$\phi_k[n] = \angle x_{a,k}[n] \tag{3.42}$$

Time-scaling is achieved through three steps. First, $\hat{a}_k$ and $\hat{\phi}_k$ are assigned using,

$$\hat{a}_k(t = \alpha n T_o) = a_k(nT_o) \tag{3.43}$$

and

$$\hat{\phi}_k(t = \alpha n T_o) = \alpha \phi_k(nT_o) \tag{3.44}$$

Secondly, the remaining samples in $\hat{a}_k$ and $\hat{\phi}_k$ are evaluated using interpolation. Finally, $\hat{x}_k[n]$ is expressed as

$$\hat{x}_{a,k}[n] = \hat{a}_k[n] cos \hat{\phi}_k[n] \tag{3.45}$$

Synthesis is achieved through the summation of all narrow band signals and re-sampling back to F_s.

Temporal smearing, transient skipping and duplication, and 'phasiness' artefacts are reduced with this method over traditional methods, and slightly improves quality over Harmonic Percussive Separation Time-Scale Modification (Sharma et al., 2017).

3.4 Source Separation Methods

Source separation methods attempt to leverage the benefits of both time and frequency-domain methods, while avoiding the detriments of both. The approach yields inconsistent results, with Harmonic Percussive Separation TSM considered a state-of-the-art method, while Non-Negative Matrix Factorisation Time-Scale Modification is not.

3.4.1 Harmonic Percussive Time-Scale Modification

Harmonic-Percussive Separation Time Scale Modification (HPTSM) of (Driedger et al., 2014) provides improvements to the quality of TSM. Time-domain methods are particularly efficient at processing percussive and noise based signals, while frequency-domain methods work best for harmonically complex signals. This difference is leveraged in HPTSM by splitting the waveform into harmonic and percussive elements, process-

ing with frequency- and time-domain methods, and then recombining. The separation process, shown in Figure 3.5, is achieved through the use of horizontal and vertical median filtering (figures 3.5a to 3.5b), creation of binary masks (figures 3.5b to 3.5c) and the application of these masks to the filtered outputs (figures 3.5c to 3.5d).

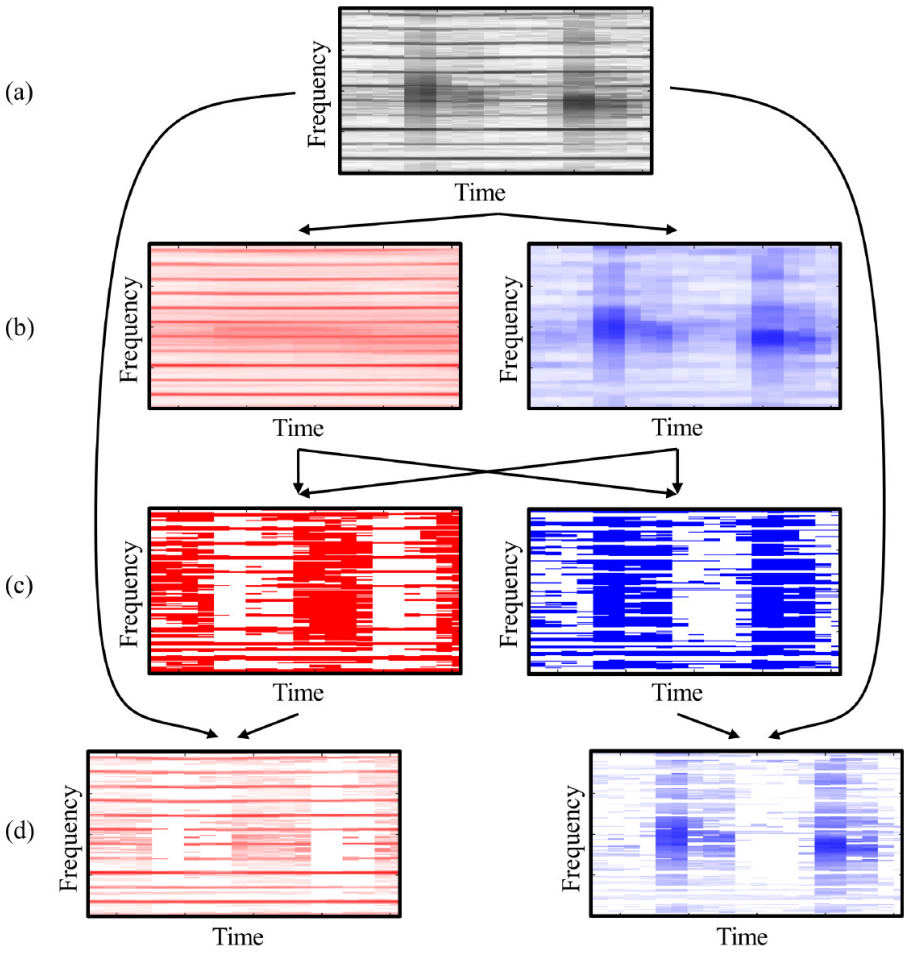

Figure 3.5: [Colour Online] Harmonic Percussive separation using binary masking. (Driedger et al., 2014) *Reprinted with permission.* ©*2013 IEEE.*

The median filter is achieved using

$$\widetilde{Y}_h(u,k) := median(Y(u-l,k),...,Y(u+l,k)) \tag{3.46}$$

and

$$\widetilde{Y}_p(u,k) := median(Y(u,k-l),...,Y(u,k+l)) \tag{3.47}$$

after computing the spectrogram of the input signal, such that $2l+1$ is the length of the filter and l is a whole number. Binary masks M_h and M_p are created in a point-wise fashion with

$$M_h := (\widetilde{Y}_h \geq \widetilde{Y}_p) \quad \text{and} \quad M_p := (\widetilde{Y}_p > \widetilde{Y}_h) \tag{3.48}$$

The harmonic and percussive spectrograms, X_h and X_p, are then calculated by applying these masks to the median filtered output using point-wise multiplication, as per

$$X_h := X \odot M_h \quad \text{and} \quad X_p := X \odot M_p \tag{3.49}$$

These spectrograms are then transformed to the time-domain by applying an inverse short-time Fourier transform (Driedger et al., 2014). The percussive signal is processed using OLA or WSOLA and the harmonic signal is processed by IPL, with the final output consisting of the sum of both time scaled outputs.

This method is effective because in a spectrogram, harmonic components occur as horizontal features, while percussive components are vertical in nature. HPTSM has been shown to be comparable to commercial TSM algorithms, and is considered the current state-of-the-art non-parametric method in published literature. A real-time implementation of this method has not yet been developed due to the large duration of signal required for the harmonic percussive separation to be effective. Ono et al. (2008) have presented a method which allows for streams of audio to be remixed with different levels of harmonic and percussive content, which may allow for a real-time implementation.

3.4.2 Non-Negative Matrix Factorisation Time-Scale Modification

Non-Negative Matrix Factorisation Time-Scale Modification (NMFTSM) of Roma et al. (2019) decomposes the signal into multiple components, each consisting of a basis function and an activation function. Analysis of the activation function allows for segmentation into sound events. These sound events are then copied to the synthesized output at the appropriate offset, and can be either transients or whole sound events. For the remaining, often harmonic, frames scaling is applied using the IPL method. The method is effective in preserving the duration of percussive events, however it is highly reliant on correct detection of the events. A feature of the method, as presented by Roma et al. (2019), is the introduction of novel TSM artefacts through extreme parameter selection.

3.5 Other Methods

3.5.1 Published Methods

A wide variety of additional methods have been published. As they are not explicitly used in the contributions of this dissertation, they have not been expanded upon. However, they have been included here for completeness. Future research would see the use of the objective measure of quality for these methods. Shape invariant time-scale modification was proposed by Quatieri and McAulay (1992); Röbel (2003) presented a strategy for transient preservation that has been widely used; fast implementation for non-linear time-scaling of stereo signals by Ravelli et al. (2005); Moinet and Dutoit (2011) proposed PVSOLA, which

is a combination of the Phase Vocoder and Synchronous Overlap-Add; Ottosen and Dörfler (2017) proposed a Phase Vocoder based on non-stationary Gabor frames; Průša and Holighaus (2017) proposed a Phase Vocoder using partial derivatives of STFT phases and their integration; and Yoneguchi and Murakami (2017) proposed a Phase Vocoder that estimates instantaneous angular frequency directly, bypassing the need for phase unwrapping.

3.5.2 Commercial Methods

A number of commercial TSM implementations are available. The most common is Elastique by Zplane Development, which is found extensively in audio software, and is considered a state-of-the-art method. Additional commercial methods include IrcamLab TS (ircamLab), Melodyne 5 (Melodyne), izotope Pitch & Time (izotope), and ZTX (zynaptiq). Free implementations include Paul Stretch (PaulStretch), which is optimised for extreme time-scaling on the order of $\alpha = 50$, and many libraries available online.

3.6 Additional Information

3.6.1 Multi-channel Considerations

The naive approach to multi-channel time scaling is to process each channel individually. However, in addition to horizontal and vertical phase coherence, channel phase coherence must also be considered. In its simplest form, stereo phase coherence ensures that there will be no change in the stereo field. This requires that the amplitude and phase relationship between channels must be maintained (Bonada, 2000). Blauert

(1997) states that partially coherent stereo signals produce larger and less sharply located stereo fields compared to a perfectly coherent signal. In frequency-domain methods the amplitude is maintained, meaning that consideration needs only be given to the phase relationship. Bonada (2002) presents a method for preserving the stereo phase coherence. By equating

$$\Delta\varphi_{analysis}(k) = \varphi_{\gamma,analysis}(k) - \varphi_{L,analysis}(k) \tag{3.50}$$

and

$$\Delta\varphi_{synthesis}(k) = \varphi_{\beta,synthesis}(k) - \varphi_{L,synthesis}(k) \tag{3.51}$$

half the difference in phase between analysis and synthesis is applied to each channel avoiding discontinuities.

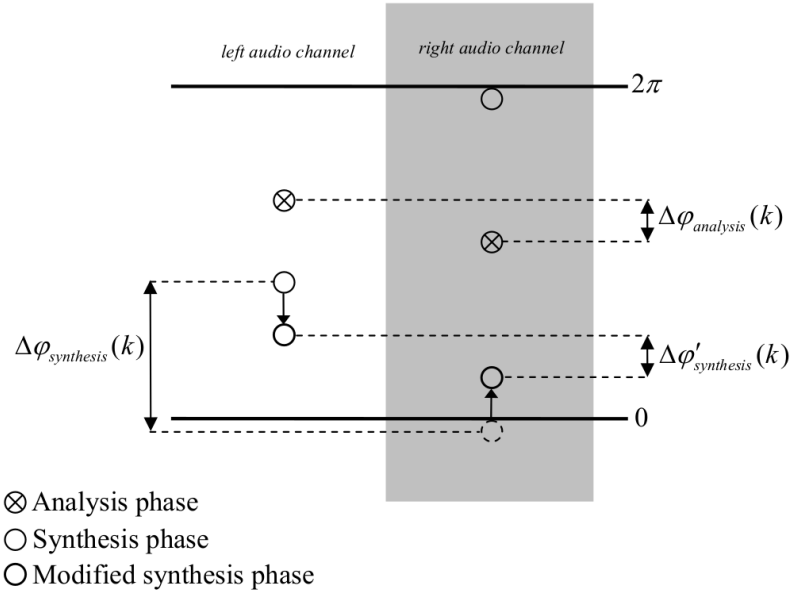

Figure 3.6: Bonada phase modification for preserving the stereo field. (Bonada, 2002). *Reprinted under CC BY-NC-ND 3.0 ES license.*

Altoe (2012) attempts to expand upon Bonada by eliminating artefacts in independent channel signals and for large stretching ratios. Altoe performs time-scaling on a single channel mix and maintains the

phase relationship between the sum and the individual channels, Figure 3.7. This method maintains the stereo phase coherence, however it can sound gritty and collapses the stereo field in some situations. It was found through implementation that equations presented by Altoe (2012) needed to be modified for the system to work correctly. The adjusted equations are given by

$$\begin{aligned} \Delta\varphi_{left}(k) &= \text{princarg}(\varphi_{mono}(k) - \varphi_{left}(k)) \\ \Delta\varphi_{right}(k) &= \text{princarg}(\varphi_{mono}(k) - \varphi_{right}(k)) \end{aligned} \tag{3.52}$$

and

$$\begin{aligned} \psi_{left}(k) &= \psi_{mono}(k) + \Delta\varphi_{left}(k) \\ \psi_{right}(k) &= \psi_{mono}(k) + \Delta\varphi_{right}(k) \end{aligned} \tag{3.53}$$

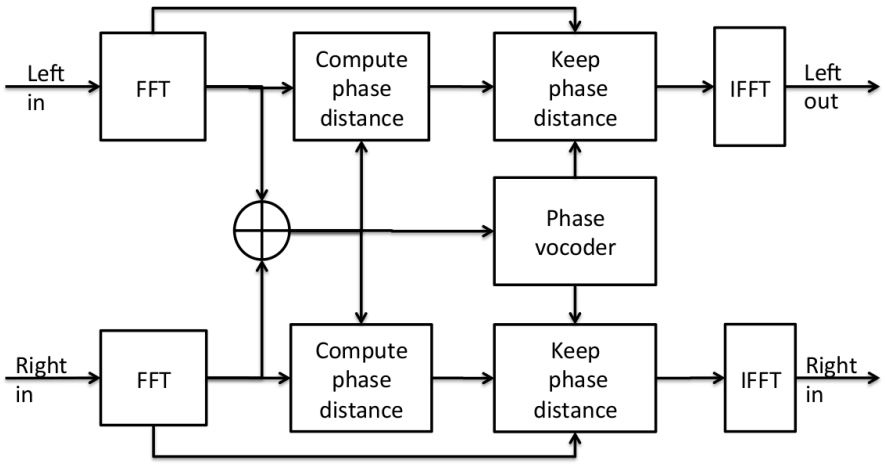

Figure 3.7: Altoe block diagram for preserving the stereo field. (Altoe, 2012). *Reprinted with permission.*

The ability to simply adjust the output phases without causing audible artefacts is presented by Dorran et al. (2004, 2005). There exists a phase flexibility within the Phase Vocoder that is dependant on the

shift size and the length of the STFT.

$$\theta = \min(0.5676, 2\text{atan}(3.6L)) \text{ radians} \tag{3.54}$$

gives the maximum phase deviation tolerated (θ) for a 50% analysis window overlap with

$$\theta = \min(0.27, 2\text{atan}(2.53L)) \text{ radians} \tag{3.55}$$

for a 75% overlap, where L is the duration of the analysis window in seconds.

Multi-channel processing for time-domain methods has not been discussed in published literature, in addition to processing more than two channels, and may provide an avenue for future research.

3.6.2 Matlab TSM library

An important tool in TSM research is the Matlab TSM Toolbox developed by Driedger and Muller (2014). This toolbox includes implementations of OLA, WSOLA, PV, IPL and HPTSM and allows for non-linear time-scaling. This allows for new algorithms and implementations to be tested against a known working method and greatly decreases research time.

Chapter 4

Measures of Quality

4.1 Introduction

Comparison of new and existing methods is integral to research, with measures of quality used to quantify the comparison. Measures of quality within signal processing are split into subjective and objective measures. Subjective testing involves participants who make subjective assessments of test material, and provides the most reliable method for assessing quality (Loizou, 2013). Subjective testing is however a time-consuming and often expensive task so efforts have been made to develop objective methods of testing, which emulate subjective testing scores. Numerous objective methods have been developed, including segmental SNR, spectral distances and perceptually motivated measures. This review will focus on Perceptual Evaluation of Speech Quality (PESQ) of Rix et al. (2001) and Perceptual Evaluation of Audio Quality (PEAQ) of Thiede et al. (2000) in addition to subjective testing methods.

Table 4.1: MOS scores for subjective assessment of sound quality or impairment ITU-T (2019)

Quality		**Impairment**	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

Table 4.2: MOS scores seven grade comparison test ITU-T (2019)

Comparison	
3	Much better
2	Better
1	Slightly better
0	The same
-1	Slightly worse
-2	Worse
-3	Much Worse

4.2 Grading Scales

The most common grading scale for subjective testing is the Mean Opinion Score (MOS) (Loizou, 2013). The International Telecommunications Union (ITU) describes an opinion score as the "value on a predefined scale that a subject assigns to his opinion of the performance of a system" (ITU-T, 2008). The MOS is simply the mean of the opinion scores across all subjects. ITU-R BS.1284-1 (ITU-T, 2019) recommends the use of three different grading scales depending on the nature and the purpose of the test, shown in tables 4.1 and 4.2.

These scales should be treated as continuous to 1 decimal point and normalisation of results for each subject is essential. For normalisation, BS.1284-1 recommends the use of

$$Z_i = \frac{x_i - x_{si}}{s_{si}}.s_s + x_s \tag{4.1}$$

where Z_i is the normalised result, x_i is the score of subject i, x_{si} is the mean score for subject i in session s, x_s is the mean score of all subjects in session s, s_s is the standard deviation for all subjects in session s and s_{si} is the standard deviation for subject i in session s.

4.3 Subjective Measures

Based on ITU-R BS.1283-1 (ITU-T, 2003) the ITU gives two possible methods for subjective testing signals without a picture and with no small impairments. These recommendations are ITU-R BS.1284-1: General methods for the subjective assessment of sound quality (ITU-T, 2019) and ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems (ITU-T, 2014).

4.3.1 General Methods

ITU-R BS.1284-1 (ITU-T, 2019) gives detailed requirements for various aspects of general subjective tests. For the subjective assessment of small impairments, more stringent requirements are recommended than will be discussed here, but can be found in ITU-T (1997). Due to the intended use for subjective testing for TSM signals, large differences in signals are involved and therefore close control of test parameters is not needed (ITU-T, 2019). For larger differences it is not essential that a reference file be used in testing. When selecting listeners for testing, expert listeners are always preferred to non-expert listeners however Streijl et al. (2016) suggests that expert listeners are more critical. ITU-T (2019) recommends a minimum of 10 expert listeners and 20 non-expert listeners and that in both cases the listeners should have training in the test procedure, materials and environment. Testing may be undertaken

using single, paired or multiple signals, all with or without a reference, with grading according to tables 4.1 and 4.2. Each set of files should not last longer than 15-20 seconds and should be presented in a random order. The total testing session should not exceed 15-20 minutes with any future sessions separated by an equal length of time. Finally if listeners are testing individually it is preferred that the listener control switching of stimuli (ITU-T, 2019). When considering sound reproduction, loudspeakers or headphones can be used provided they provide neutral sound and that the prominence of certain artefacts for each reproduction method is considered. The use of headphones can reduce variability as reproduction is independent of the room in which the testing is undertaken (ITU-T, 2019).

4.3.2 MUSHRA

MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) is a subjective testing method described in ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems (ITU-T, 2014). MUSHRA is suited to evaluating the subjective quality of lower quality audio systems and shares many recommendations with ITU-R BS.1284-1 ITU-T (2019) and ITU-R BS.1116-1 ITU-T (1997). MUSHRA testing uses sets of signals which include a low-pass filtered version of the reference signal known as the anchor. By using multiple stimulus, a known reference, a hidden reference and a hidden anchor MUSHRA is considered appropriate for medium and large impairments (Soulodre and Lavoie, 1999). The ability to switch stimuli at will gives the listener the ability to compare multiple each stimulus to any other stimulus. The number of stimuli should be limited to no more than 15 signals in any trial (ITU-T, 2014). Experience with the system has

shown that listeners begin each trial with a rough estimation of quality, followed by sorting, and finally grading of quality (ITU-T, 2014). Each stimuli is scored on a continuous quality scale divided into 5 equal intervals using the sound quality labels from Table 4.1. A slider for each stimulus should be displayed with the ability to adjust the slider while listening to that stimulus. More information about interface design can be found in ITU-T (2014).

4.3.3 Subjective Testing of Time-Scale Modification

Comparatively little formal subjective testing has been used to evaluate proposed methods, with most proposed methods providing results from informal testing. Papers to include formal subjective testing include Karrer et al. (2006), Jun et al. (2007), Moinet and Dutoit (2011), Driedger et al. (2014), Damskägg and Välimäki (2017) and Sharma et al. (2017), while informal subjective testing was reported by many papers, including Roucos and Wilgus (1985), Verhelst and Roelands (1993), Puckette (1995), Laroche and Dolson (1999), Bonada (2000), Dorran et al. (2004), Kafentzis et al. (2013) and Roma et al. (2019). A wide variety of time-scales and algorithms are used, with little consistency. Time-scales are often limited with two to five times scales ($0.5 \leq \beta \leq 2$) reported in formal testing, with a bias towards $\beta < 1$. This reduces the number of files that require rating, but also limits algorithm evaluation. The difference in quality between $\beta < 1$ and $\beta > 1$ was mentioned briefly by Sylvestre and Kabal (1992) and shown in early results from this testing in (Roberts and Paliwal, 2019). Since the release of the MATLAB TSM Toolbox (Driedger and Muller, 2014), the included algorithms, PV, IPL, WSOLA and HPTSM, have been used in most evaluations, while comparisons to commercial algorithms are rare (Kar-

rer et al., 2006; Driedger et al., 2014; Damskägg and Välimäki, 2017). The source audio used during testing also varies between papers with some papers using the files provided with the MATLAB TSM Toolbox. It was noted by Moulines and Laroche (1995) that a thorough perceptual evaluation of TSM approaches and algorithms had not yet been undertaken.

4.4 Objective Measures

Objective measures of quality are developed to predict results from subjective testing with a high level of accuracy (Loizou, 2013) with systems incorporating properties of the human auditory system existing since 1979 (Thiede et al., 2000). By modelling the response of the ear and cognition systems a more accurate system can be designed. PESQ and PEAQ are both perceptually motivated objective measures of quality. The resurgence of machine learning has seen a marked increase in objective measures of quality for speech, including "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech" of Patton et al. (2016) and "Non-intrusive speech quality assessment using neural networks" of Avila et al. (2019).

4.4.1 Traditional Time-Scale Modification Measures

Only two measures of quality have been proposed for TSM. These are Signal to Error Ratio (SER) of Roucos and Wilgus (1985) and synthesis consistency ($\mathrm{D_M}$) of Laroche and Dolson (1999). Both of these measures draw from the work of Griffin and Lim (1984), however neither have seen continued use. Laroche and Dolson (1999) note that $\mathrm{D_M}$ provides only a high level indicator of 'phasiness'. There has been no comparison

between subjective evaluations and either measure.

Signal to Error Ratio

Signal to Error Ratio (SER) was introduced by Griffin and Lim (1984) to study the convergence of their LSEE-MSTFTM algorithm. Roucos and Wilgus (1985) used the Signal to Error Ratio (SER),

$$SER = 10 \log_{10} \frac{\sum_{u=0}^{\mathrm{U}-1} \sum_{k=0}^{\frac{N}{2}} |X_T(u,k)|^2}{\sum_{u=0}^{\mathrm{U}-1} \sum_{k=0}^{\frac{N}{2}} (|X_R(u,k)| - |X_T(u,k)|)^2} \tag{4.2}$$

to compare the quality of iterative TSM methods, but found that SER could only be regarded as a general measure of quality, with signals of a similar SER resulting in different subjective quality evaluations. This is because SER only accounts "for the magnitudes of the successive spectra, and does not directly concern their phases" (Roucos and Wilgus, 1985). u is the frame number, k is the frequency bin, U is the total number of frames, N is the frame size, X_R is the Short Time Fourier Transform (STFT) of the reference signal and X_T is the STFT of the test signal.

Synthesis Consistency

Laroche and Dolson (1999) proposed a method of calculating the phase consistency of the output,

$$\mathrm{D_M} = \frac{\sum_{u=0}^{\mathrm{U}-1} \sum_{k=0}^{\frac{N}{2}} (|X_T(u,k)| - |X_R(u,k)|)^2}{\sum_{u=0}^{\mathrm{U}-1} \sum_{k=0}^{\frac{N}{2}} |X_R(u,k)|^2} \tag{4.3}$$

as a way of validating the vertical phase coherence of the synthesised signal. This measure of consistency was developed because for complex signals, no simple phase relationship within the sound exists and an

analytical formula was difficult to develop (Laroche and Dolson, 1999). To avoid taking into account errors due to missing overlapped segments in the resynthesis formula a number of frames from the start and end of the signal are excluded from the calculation (Laroche and Dolson, 1999). If total vertical phase coherence is achieved, D_M will equal 0.

Laroche and Dolson found the above measure of consistency didn't "always accurately reflect the perceived 'phasiness' of the modified signal" Laroche and Dolson (1999). A number of examples are given with the conclusion that D_M doesn't clearly indicate 'phasiness'.

4.4.2 Signal to Noise Ratio

A common method for determining the quality of a signal is the Signal to Noise Ratio (SNR). Expressed in decibels, the measure is the ratio between the energy of the signal and that of the noise. It is calculated according to

$$SNR = \frac{\boldsymbol{\sigma}_x^2}{\boldsymbol{\sigma}_e^2} = \frac{E(x^2(n))}{E(e^2(n))} \tag{4.4}$$

from Huang et al. (2001), where $E(\cdot)$ is the expectation or mean of the signal and e denotes the error signal.

SNR can however give a poor indication of quality for non-stationary signals, such as speech. Consequently, segmental SNR is often used (Thampi et al., 2014). It is calculated using

$$SNR_{seg} = \frac{10}{U} \sum_{u=0}^{U-1} \frac{\sum_K^{k=1} W(k,u) log_{10} \frac{|X(k,u)|^2}{(X(k,u) - |\hat{X}(k,u)|^2)^2}}{\sum_K^{k=1} W(k,u)} \tag{4.5}$$

where $W(k,u)$ is the weight of the k^{th} frequency band, U is the total number of frames in the signal, K is the number of banks. $X(k,u)$ and $\hat{X}(k,u)$ are the weighted clean signal spectrum and weighted enhanced

signal spectrum respectively in the k^{th} frequency band at the u^{th} frame.

4.4.3 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ), ITU-R P.862 (ITU-T, 2001b), is a perceptually motivated objective measure of quality for speech within a wide range of network conditions Rix et al. (2001). The structure of PESQ can be seen in Figure 4.1 and can be generalised to pre-processing, time alignment, auditory transform, disturbance computation and MOS prediction.

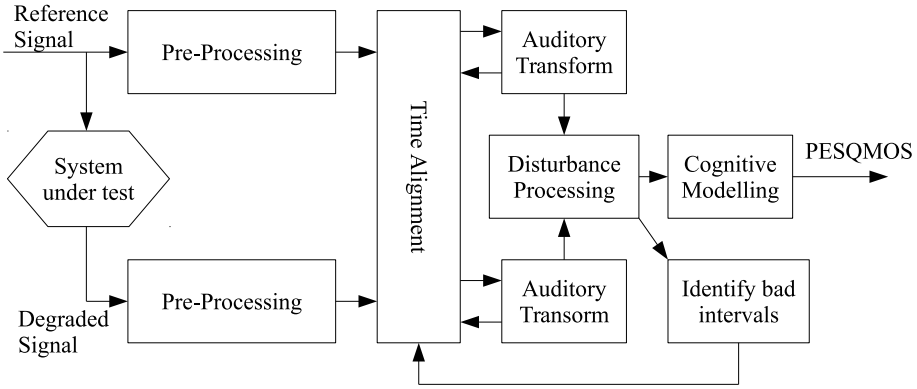

Figure 4.1: PESQ Block Diagram. (Loizou, 2013). *Reprinted with permission.*

Pre-processing aligns the level of the input signals to standard listening levels using gains based on RMS levels for band-pass (350-3250 Hz) filtered speech. The signals are then time-aligned using two stages. Crude delay estimation using cross-correlation between the two signals, is followed by division of the reference file into utterances, which are then then aligned using cross-correlation. The psycho-acoustic model of PESQ maps signals to a representation of perceived loudness in time and frequency. The instantaneous power spectrum is grouped without smearing into 42 bins equally spaced on the Bark scale and is followed

by frequency equalisation between the reference and degraded signals (Rix et al., 2001). Gain equalisation is also calculated to compensate for short-term gain variations. The ratio between audible power for reference and test signals is used for the gain equalisation (Loizou, 2013). Finally, the Bark scale is mapped to Sone loudness giving perceived loudness for each frame, known as loudness densities. Raw disturbance values are calculated using the signed difference between the degraded and original loudness densities. When components such as noise have been added raw disturbance values are positive, while negative values are a result of omitted components (ITU-T, 2001b). Minima from the loudness densities are used to create a mask to model the inaudible nature of small differences. After masking, the result is known as symmetric disturbance. Asymmetric disturbances are also calculated to account for disturbances introduced by codecs, and are calculated by multiplying the symmetric disturbance by an asymmetric disturbance factor per frame. The symmetric and asymmetric disturbances are aggregated using different L_p norms and emphasising frames with low loudness. These values, d_{SYM} and d_{ASYM}, are used when calculating the predicted MOS score. MOS score prediction is calculated using

$$MOS_{PESQ} = 4.5 - 0.1d_{SYM} - 0.0309d_{ASYM} \tag{4.6}$$

and ranges from 1.0 (bad) to 4.5 (no distortion). For extremely high distortion PESQMOS may fall below 1.0, however this is uncommon (Rix et al., 2001)

4.4.4 Composite Speech Quality Measures

In addition to PESQ, composite quality measures have been proposed (Hu and Loizou, 2007). Three common measures are the Composite Signal Distortion (C_{sig}), Composite Background Noise Distortion (C_{bak}) and Composite Overall Quality (C_{ovl}) (Thampi et al., 2014), and are calculated using

$$C_{ovl} = 1.594 + 0.805MOS_{PESQ} - 0.007WSS - 0.512LLR \tag{4.7}$$

$$C_{sig} = 3.093 + 0.603MOS_{PESQ} - 0.009WSS - 1.029LLR \tag{4.8}$$

and

$$C_{bak} = 1.634 + 0.478MOS_{PESQ} - 0.007WSS + 0.063SNR_{seg} \tag{4.9}$$

LLR is the Log Likelihood ratio, given by

$$LLR(\overrightarrow{a}_p, \overrightarrow{a}_c) = log\left(\frac{\overrightarrow{a}_p \boldsymbol{R}_c \overrightarrow{a}_p^T}{\overrightarrow{a}_c \boldsymbol{R}_c \overrightarrow{a}_c^T}\right) \tag{4.10}$$

where $\overrightarrow{a}_c$ and $\overrightarrow{a}_p$ are Linear Prediction Coefficients for original and enhanced speech, and $\boldsymbol{R}_c$ is the auto-correlation matrix of the original speech. WWS is the Weighted Spectral Slope, given by

$$WSS = \frac{1}{U}\sum_{m=0}^{U-1}\frac{\sum_K^{k=1} W(k,u)(S_c(k,u) - S_p(k,u))^2}{\sum_K^{k=1} W(k,u)} \tag{4.11}$$

where $W(k,u)$ are weights as per Klatt (1982) and $S_c(k,u)$ and $S_p(k,u)$ are spectral slopes. Spectral slopes are calculated as the difference between adjacent spectral magnitudes in decibels.

4.4.5 PEAQ

Perceptual Evaluation of Audio Quality (PEAQ), ITU-R BS.1387-1 (ITU-T, 2001a), is a perceptually motivated objective measure of quality, developed primarily for evaluation of audio codecs. Released as an ITU standard in 2001, it combines seven different proposed methods, namely DIX, NMR, OASE, PAQM, PERCEVAL, POM and Toolbox (Thiede et al., 2000). Figure 4.2 shows a block diagram of the PEAQ method based on (ITU-T, 2001a). PEAQ consists of peripheral ear models, calculation of mostly psychoacoustic Model Output Variables (MOVs) and mapping to a single value using a neural network (ITU-T, 2001a). Several intermediary steps of pre-processing excitation patterns are also included.

PEAQ has two modes of operation, basic and advanced. The basic version (PEAQB) consists of a Fast Fourier Transform (FFT)-based peripheral ear model and 11 Model Output Variables (MOVs). It aims to process the input signals in a similar manner to the ear. As such, absolute threshold, frequency scaling, basilar excitation, discrimination between signals, masking, loudness and partial loudness are all considered in the peripheral ear models. The basic model begins with an FFT, followed by rectification, scaling of the input signal, outer and middle ear weighting, mapping into auditory filter bands, addition of internal noise, frequency-domain spreading and time-domain spreading. The advanced version (PEAQA) follows the same framework, but with a filter-bank-based ear model and 5 MOVs. It contains scaling of the input signals, DC rejection, auditory filter band decomposition, outer and middle ear weighting, frequency-domain spreading, rectification, time-domain spreading, adding of internal noise and additional time-domain spreading. Pre-processing of the resulting excitation patterns for both

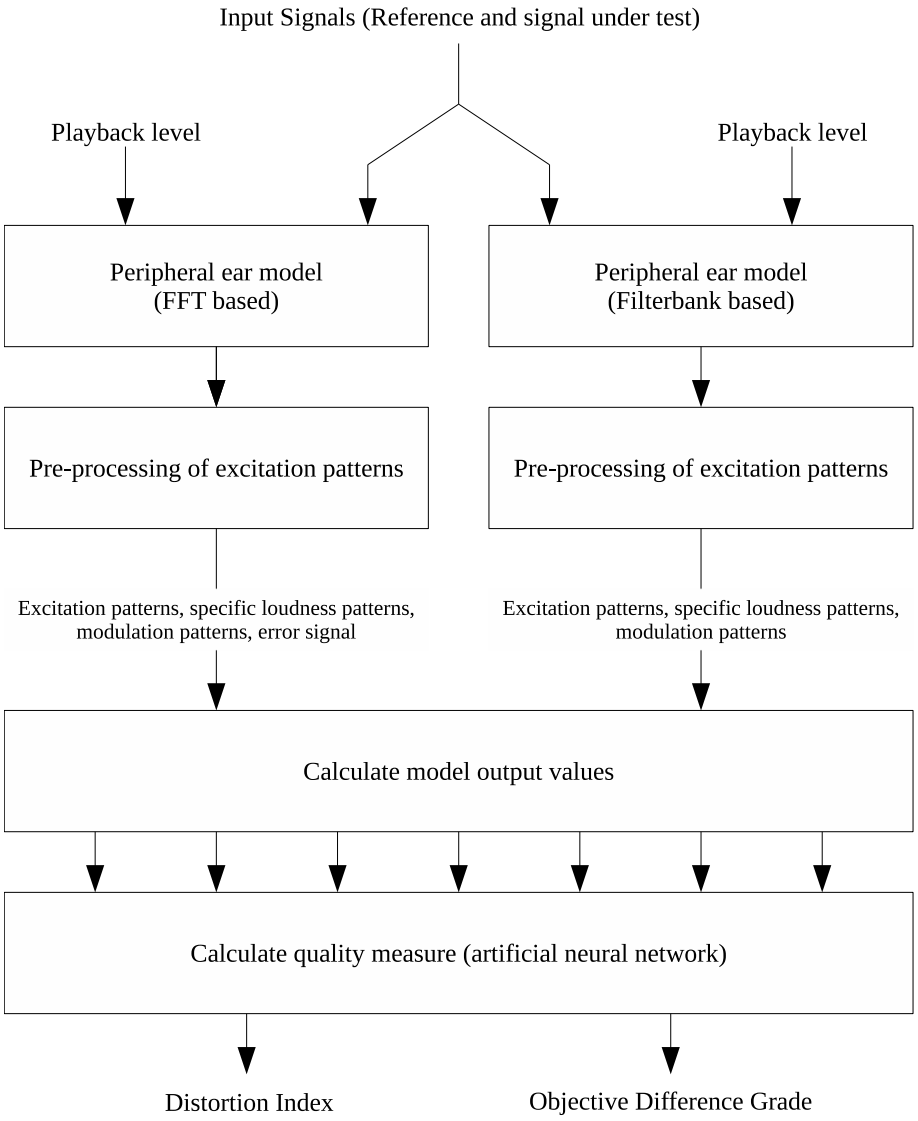

Figure 4.2: PEAQ Block Diagram. (ITU-T, 2001a). *Reprinted with permission.*

ear models creates patterns used in the calculation of the MOVs, the details of which can be found in Thiede et al. (2000), ITU-T (2001a) and Kabal et al. (2002).

The basic MOVs can be categorised into six groups. Modulation difference MOVs, *WinModDiff1B*, *AvgModDiff1B* and *AvgModDiff2B*, are

the windowed and linear averages of the modulation differences. Noise loudness MOVs, of which *RmsNoiseLoudB* is the only one used in the basic method, is the squared average of the noise loudness and takes masking into account. Bandwidth MOVs, *BandwidthRefB* and *BandwidthTestB*, estimate the mean bandwidth of the reference and testing signals considering only frames with a bandwidth greater than 8kHz. Psuedocode for the calculation is given in ITU-T (2001a). When considering auditory masking, *Total NMRB*, is the linear mean of the noise-to-mask ratio, while Relative Disturbed Frames Basic, *RelDistFramesB*, is the number of frames with a noise-to-mask ratio above 1.5dB as a ratio of the number of frames for the signal. For detection probability, Maximum Filtered Probability of Detection (*MFPDB*) models the smaller impact of distortions at the beginning of the file on quality assessment. Average Distorted Block (*ADBB*), uses the number of frames with a distortion detection probably above 0.5 and is calculated according to Section 4.7.2 in ITU-T (2001a). Finally, the Harmonic Structure of Error (*EHSB*) MOV measures the harmonic structure of the error signal, as strong harmonic structure may be transferred to the error signal. The advanced model adds an additional four MOVs, while also using *EHSB*. *RmsModDiffA*, *RmsNoiseLoudAsymA* and *AvgLinDistA* are all calculated from the filterbank ear model excitation patterns, while *SegmentalNMRB* is calculated from the FFT model. For full details, see ITU-T (2001a) and Kabal et al. (2002).

While it is difficult to specify the mapping from peripheral to cognitive representation, PEAQ considers prior knowledge, learning, the difference between opinion of additive and subtractive time-frequency components, the nature of distortion and non-uniform weighting of sections within test files. PEAQ makes use of a neural network to map the

MOVs to a single Distortion Index (DI) value. The PEAQ basic version neural network structure has 11 inputs, 1 hidden layer with 3 nodes and 1 output node while the advanced version has 5 inputs, 1 hidden layer with 5 nodes and 1 output node (ITU-T, 2001a). Features are scaled to between 0 and 1, using

$$\hat{MOV}[i] = \frac{MOV[i] - a_{min}[i]}{a_{max}[i] - a_{min}[i]} \tag{4.12}$$

where a_{min} and a_{max} are scaling factors, before input to the network.

The sigmoid activation function as well as node weights are detailed in ITU-T (2001a). Finally, the DI is mapped to the final Objective Difference Grade (ODG) using

$$ODG = b_{min} + (b_{max}[i] - b_{min}[i]) \cdot sig(DI) \tag{4.13}$$

where b_{min} and b_{max} are selected to minimise the Root Mean Square Error (RMSE).

During validation PEAQA achieved a correlation of $r = 0.851$ for an untested database and resulted in less outliers than other methods during testing (ITU-T, 2001a).

For detailed information, the reader is directed to ITU-T (2001a) and Kabal et al. (2002). The initial PEAQ standard was found to contain errors and omit vital information required for a proper implementation of the standard. Kabal et al. (2002) clarified these errors and omissions, and provided a MATLAB implementation of the standard. Available implementations include PQeval (Kabal et al., 2002), gstpeaq (Holters, 2017), EAQUAL (godock, 2017), peaqb-fast (Gottardi, 2013) and PEAQPython (Welch and Cohen, 2015).

4.4.6 Related Objective Measures of Quality

A number of related objective measures have been proposed, including PEMO-Q of Huber and Kollmeier (2006), and the 2f-model of Kastner and Herre (2019) that makes use of two PEAQ Basic features.

The objective measure of Huber and Kollmeier (2006), uses the perceptual model of the auditory system of Dau et al. (1997) to simulate the transformations from acoustic to neural signals made by the ear. It is an expansion of the speech quality measure of Hansen and Kollmeier (2000) and transforms the signal using basilar-membrane filtering, half-wave rectification, low-pass filtering, absolute thresholding, adaptation and modulation filtering. The linear cross-correlation between transformed reference and test signals gives the perceptual quality measure for each channel of the basilar-membrane filterbank. A weighted sum of the resulting correlation coefficients is then used to calculate the final perceptual quality measure (PSM). An additional measure is calculated by computing successive cross-correlations for 10ms frames of the internal representation. This signal is weighted by the internal representation of the test signal's moving average. Finally, the fifth percentile of the weighted time-series is used as an additional quality measure (PSM_t). Huber and Kollmeier (2006) showed the second measure to be independent of the type of input signal. Performance is slightly better than PEAQ for known signals (PEMO-Q: $r = 0.9$, PEAQA: $r = 0.87$, PEAQB: $r = 0.89$) and significantly better for unknown signals (PSM: $r = 0.97$, PSM_t: $r = 0.88$, PEAQA: $r = 0.79$). However, if speech and music signals are considered independently, PEAQA results improve to $r = 0.96$ for speech and $r = 0.98$ for music. In the context of linear distortions, PEAQ achieves better results than PEMO-Q (PSM: $r = 0.69$, PSM_t: $r = 0.66$, PEAQA: $r = 0.79$), likely due to PEAQ explicitly

accounting for linear distortions. When considering computational complexity, PEMO-Q is approximately 12 times higher than PEAQ Advanced (Huber and Kollmeier, 2006).

A simplification of PEAQ was proposed by Kastner and Herre (2019). This model efficiently estimates the subjective quality of source separated signals. The method uses the *AvgModDiff1B* and *ADBB* features in

$$\mathrm{MMS}_{est} = \frac{49.73}{1 + (-0.0315 AvgModDiff1B - 0.783)^2} - 46.96 ADBB + 147.12 \tag{4.14}$$

It achieves an overall accuracy of 0.81 and outperforms other state-of-the-art objective measures for source separated signals including BSSEval of Vincent et al. (2006), PEASS of Emiya et al. (2011), PEMO-Q, PEAQ and the four feature model of (Kastner, 2009).

4.4.7 Figures of Merit

With every objective measure of quality there must be a method for determining the merit of the measure. Loizou (2013) suggests the creation of a large database of signals distorted by various means and evaluate each signal via subjective testing. Statistical analysis can then be used to calculate the correlation between the objective and subjective measures of quality. The Pearson correlation coefficient,

$$P_k = \overline{P} + \rho \frac{\sigma_P}{\sigma_O}(O_k - \overline{O}) \tag{4.15}$$

where O_k is the k^{th} objective score, P_k is the k^{th} subjective score, σ_P and σ_O are the standard deviations of subjective and objective scores respectively, and $\overline{P}$ and $\overline{O}$ are the mean subjective and objective scores

respectively, can be used for this purpose.

A second figure of merit is the standard error of the estimate and is calculated using

$$\sigma_e = \sigma_P \sqrt{1 - \rho^2} \tag{4.16}$$

and estimates the standard deviation of the objective measure error (Loizou, 2013). These two values provide a useful measure of the merit of an objective measure.

Part III

Research

Chapter 5

Time-Scale Modification Improvements

5.1 Stereo Time-Scale Modification

5.1.1 Introduction

Most published TSM methods ignore application in multi-channel environments. Exceptions to this are Bonada (2000), Ravelli et al. (2005) and Altoe (2012), however the stereo field is considered an after-thought, with no published results on the improvement made through the proposed algorithm. Bonada (2000) presents that the phase relationship between each of the signals must be considered when processing multi-channel signals and proposes post-TSM phase adjustment to maintain the stereo channel phase relationship. This method is effective; however, it suffers at slow timescales and when processing independent channels. Ravelli et al. (2005) uses cross-correlation at transient onsets to align transients in each channel, increasing channel phase coherence. This

stereo method is only applicable to the specific TSM method described in the paper, and does not consider the phase relationship for non-transient content. Instead, it aims to increase the phase coherence rather than maintain the original relationship. Finally, Altoe (2012) attempts to improve on Bonada (2000) by processing the sum of the channels and maintaining the phase relationship between the sum and independent channels. The methods presented by Bonada and Altoe are however constrained to frequency-domain methods due to modifying the phase spectra of each frame during TSM.

Blauert (1997) states that partially coherent stereo signals produce larger and less sharply located stereo fields in comparison to a perfectly coherent signal. If the channels of a multi-channel source are processed independently, the phase relationship between the signals can be lost resulting in a distorted stereo field. Small time delays between channels for time-domain methods and changes in phase spectra for frequency-domain methods cause a change in phase relationship between channels. This change in phase relationship is perceived as the sound source moving to the outside of the stereo field. This effect only occurs when there are differences between the channels, such as a stereo recording with natural reverberation. As described in Section 2.2, sum and difference is a useful stereo signal representation that finds use in applications from stereo FM radio transmission to the Mid-Side microphone technique. The representation is mono-compatible with the sum containing in-phase signal information, while the difference contains the out-of-phase signal information. It should be noted that while the sum signal is generally of a greater energy level (4-6 dB), compared to the difference signal, this is of little importance to the proposed method as the signal is transformed to and from the sum and difference representation.

The method proposed in this chapter utilises the sum and difference representation of a two channel signal to improve the quality of the time-scaled signal. Two methods are presented with the first transforming the entire signal prior to and after TSM, and the second transforming the signal after framing and then prior to overlap-add reconstruction. Objective and subjective testing is undertaken comparing the proposed method with Bonada (2000), Altoe (2012), and the naive method of processing each channel independently.

The proposed method is presented in Section 5.1.2, testing methodology is presented in Section 5.1.3, results are presented in Section 5.1.4, and the conclusion is presented in Section 5.1.5.

5.1.2 Method

The proposed method uses sum and difference signals during TSM, which allows for implementation within or external to existing TSM methods, illustrated in Figure 5.1. Pre-processing creates the sum and difference signals, while post-processing converts back to the left and right representation.

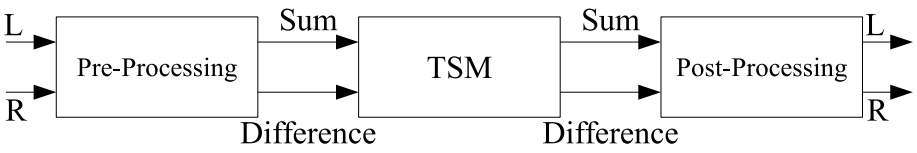

Figure 5.1: Block diagram for the proposed file method. Pre-processing transforms the signal to the sum and difference representation, while post-processing transforms the scaled signal back to left and right.

During pre-processing the stereo signal is transformed using

$$S_1 = L + R \quad ; \quad D_1 = L - R \tag{5.1}$$

where S_1 is the element-wise sum of left and right samples, and D_1 is the element-wise difference between left and right samples. This process is reversible without loss using

$$L_1 = \frac{S_1' + D_1'}{2} \quad ; \quad R_1 = \frac{S_1' - D_1'}{2} \tag{5.2}$$

Post-processing transforms the modified sum and difference signals back to the standard left and right signals, using Equation 5.2. In its simplest form this transform is applied to the entire audio file, but can also be applied to each frame in a real-time application. Due to the orthogonal nature of the sum and difference representation, any change moves the resultant phase relationship towards phase coherence, resulting in a processed signal that does not lose presence in the centre of the stereo field.

It was found empirically that the system is improved in some cases if Equation 5.1 is used for signals biased to the left and

$$S_2 = L + R \quad ; \quad D_2 = R - L \tag{5.3}$$

is used for signals biased to the right. As a result, an additional frame based method was developed. The method first calculates the peak level balance of the stereo field, using

$$\hat{B}(n) = \frac{|x_L(n)|}{max[|x|]} - \frac{|x_R(n)|}{max[|x|]} \tag{5.4}$$

and

$$B = \frac{1}{N} \sum_{n=0}^{N-1} \hat{B}(n) \tag{5.5}$$

to determine the appropriate set of equations for calculating the sum and difference signals.

Equations 5.1 and 5.2 are used for left biased signals, while equations 5.3 and

$$L_2 = \frac{(S'_2 - D'_2)}{2} \quad ; \quad R_2 = \frac{(S'_2 + D'_2)}{2} \tag{5.6}$$

are used for right biased signals. This method allows for the signal balance to shift between channels. This modification also preserves the balance when processing signals with silence in either channel. However, due to the use of overlapping frames in many TSM methods, this method must be implemented within the TSM algorithm. Alternatively, if the entire signal is framed before processing, the bias may be calculated before time scale processing.

5.1.3 Testing

Objective and subjective methods were used during evaluation. To facilitate objective testing two features were developed, similar to that found in Avendano and Jot (2002) and Ravelli et al. (2005). The two features developed, Stereo Phase Coherence (SPC) and Balance, give an indication of important features within the stereo field. As the name suggests, SPC is used to measure the phase coherence between each channel and is calculated in the time-domain. SPC also gives an indication of the perceived width of the stereo field. Balance is used to measure the mid-point, pan or bias of the signal. These features were chosen due to their use in Ravelli et al. (2005) and extensive use in music production software to give visual feedback about the signal under examination.

The SPC feature, C, shows the average time-domain phase coherence of the frame under investigation and ranges from 1 (completely coherent) to -1 (completely incoherent). It is calculated by framing the signal and

normalising the element wise multiplication between the channels,

$$\hat{C}(n) = \frac{x_L(n)x_R(n)}{max[|x|]} \tag{5.7}$$

Each value is subsequently bounded such that positive values are set to 1 (in-phase) and negative values are set to -1 (out-of-phase), as per

$$\hat{C}_{SIGN}(n) = \begin{cases} 1 & \hat{C}(n) > 0 \\ -1 & \hat{C}(n) < 0 \\ 0 & \hat{C}(n) = 0 \end{cases} \tag{5.8}$$

The phase coherence for each frame is finally calculated as the mean of each frame,

$$C = \frac{1}{N} \sum_{n=0}^{N-1} \hat{C}_{SIGN}(n) \tag{5.9}$$

The frames are subsequently concatenated to form the feature. By using this method, many cross-correlations are removed from the method of Ravelli et al. (2005).

The balance feature, B, shows the midpoint of the stereo field for the frame under investigation and ranges from 1 (Left) to -1 (Right). It is calculated as the difference between the normalised absolute values of the left and right channels, shown above in Equation 5.4. The mean of each frame is then calculated, as in Equation 5.5, with the result for each frame concatenated to form the feature.

Examples of these features for synthetic test files can be seen in Figure 5.2. *White SD Fade* is a signal containing two independent channels of white noise fading between sum and difference representations. As can be seen, a signal fading from sum to difference maintains a balance in the centre of the stereo field, while the coherence moves from in-phase

to out-of-phase. *Sine panning* is a sine tone panning from right to left. The resulting coherent signal moves maintains a coherence of 1, while the balance moves from negative to positive. This same test for a single DC channel results in balance at its full range.

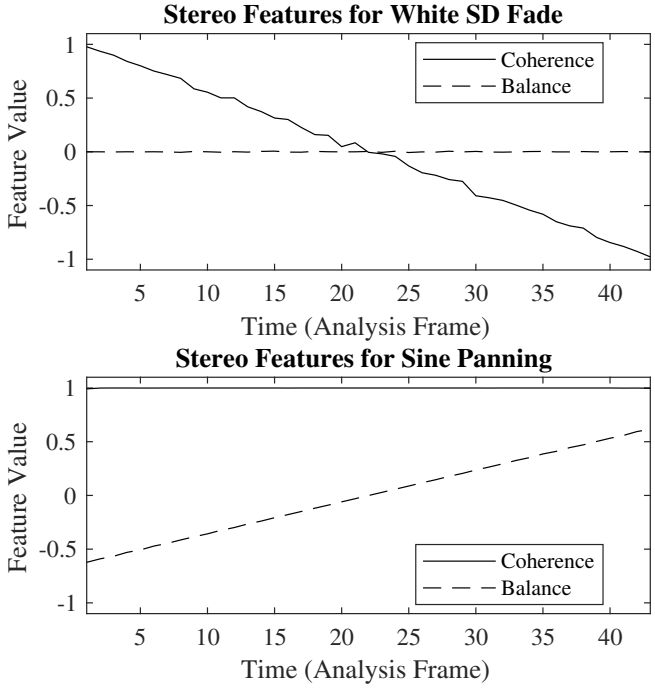

Figure 5.2: Stereo features for white noise cross-fading in the sum and difference representation and a sine tone panning right to left.

During evaluation, each of the features were calculated from both the reference and test signals. The reference features were linearly interpolated to match the length of the test signal features, before L2 norms,

$$\hat{B}_D = \sqrt{(B_x(u) - B_y(u))^2} \tag{5.10}$$

and

$$\hat{C}_D = \sqrt{(C_x(u) - C_y(u))^2} \tag{5.11}$$

were used to calculate the distance between the processed and inter-

polated signals. $\hat{B}_D$ and $\hat{C}_D$ are the feature distances, u is the frame number, $B_x(u)$ and $B_y(u)$ are balance features for the reference and test signals, and $C_x(u)$ and $C_y(u)$ are SPC features for the reference and test signals. The resulting distances were then averaged, using

$$B_D = \frac{1}{U} \sum_{u=0}^{U-1} \hat{B}_D(u) \tag{5.12}$$

and

$$C_D = \frac{1}{U} \sum_{u=0}^{U-1} \hat{C}_D(u) \tag{5.13}$$

for each feature resulting in a dissimilarity for each of the TSM methods under testing. B_D and C_D are the mean feature distances and U is the total number of frames.

Double-blind subjective testing was undertaken with twelve participants evaluating four sets of files processed with Naive, Proposed, Bonada and Altoe Phase Vocoder implementations at $\beta = 0.8258$. Participants, with backgrounds in signal processing and music technology, were trained using an additional set of files that portray a change in balance, a comparison of stereo width and the loss of phase coherence. The participants were then played the reference source material for familiarisation before testing. Each evaluation consisted of the playback of the reference file followed by a pair of test files. Participants were asked to select the file that had the highest similarity to the stereo field of the reference file and were asked to pay specific attention to the location of sound sources within the stereo field. One point was given to the chosen method, with points split evenly between methods if the test files were judged to sound the same. All permutations for each set were presented in random order resulting in 48 signal pairs for each participant, and 576 total evaluations. Sound reproduction was through Sennheisser HD280

headphones, in a quiet office, with files normalised before playback.

For objective testing, eleven TSM ratios (0.3838, 0.4427, 0.5383, 0.6524, 0.7821, 0.8258, 0.9612, 1.257, 1.4692, 1.6961 and 1.8412), were applied to the reference audio files listed in Table 5.1. Naive, Altoe, Bonada, Both Proposed methods were used to process the reference files. Modified versions of WSOLA and HPTSM methods were used with naive and proposed stereo processing, resulting in 396 test files. Features for each of these files were calculated followed by the mean dissimilarity calculation. Audio files all had a sample rate of 44.1kHz and a bit depth of 16 bits. Features were extracted using a frame length of 2048 samples (42.4ms). Non-integer TSM ratios were not used to ensure a loss of phase coherency due to phase unwrapping errors (Laroche and Dolson, 1999). WSOLA and HPTSM methods were implemented around the MATLAB TSM Toolbox (Driedger and Muller, 2014), with new MATLAB implementations for all other methods. The frame-based sum and difference method was implemented in a traditional vocoder such that both channels were processed simultaneously, giving the ability to produce sum and difference signals for each frame. The proposed method was also implemented within the Extempore programming environment (Sorensen and Gardner, 2017).

Table 5.1: Reference audio files with description.

Test File	**Comments**
Choral	Choir in a small reverberant church.
Electropop	Synthetic polyphonic music with specific panning of sounds, to test reproduction of stereo features and moving sound sources.
Jazz	Big Band with vocalist. Strong central voice and percussion elements, with wind instruments panned through the stereo field.
Saxophone Quartet	Wide stereo field, low reverberation.

To further examine the resulting features, particularly the large variance in dissimilarity, a larger dataset of 88 files was processed, resulting in 8712 test files. This dataset will be fully detailed in Section 6.3. The overall SPC and balance for processed and original files was graphed. These reference files, contain multi-mono signals, stereo recordings of a single source and complex stereo fields. This allowed for the impact of the original stereo field, the TSM ratio and the sound source to be evaluated. A smaller subset of 12 files was used to generate plots for this chapter to increase clarity.

5.1.4 Results

Subjective testing, figure 5.3, shows a clear preference for Altoe, Bonada and the Proposed method over the naive method when considering the representation of the stereo field. The proposed method was judged to be comparable to Altoe and approaching Bonada for the files evaluated. Expert listeners were able to more accurately detect changes in the stereo field with results further in favour of the proposed method, with all participants reporting that it was often difficult to discriminate between the test material. Participants also reported that the proposed method maintained the centre presence of the signal, however there was a slight narrowing of the stereo field in some test cases.

The proposed whole-file and frame methods show a large improvement over the naive approach of single channel implementations for the SPC feature, as per figure 5.4. The general nature of the proposed method can also be seen with reduced dissimilarity for frequency-domain and time-domain methods. A small increase in dissimilarity for the balance feature is observed for the proposed methods, figure 5.5, however this change is negligible. The small dissimilarity for all methods con-

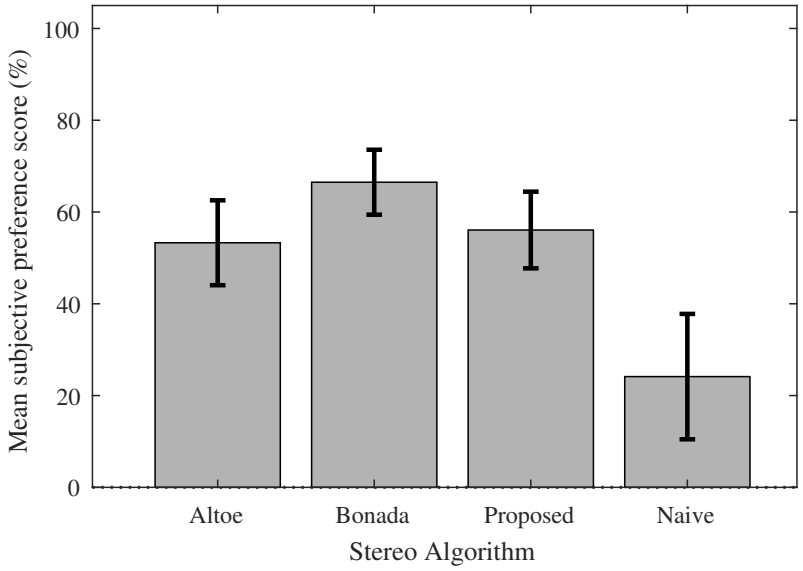

Figure 5.3: Mean Subjective Preference score for Altoe, Bonada, Proposed and Naive Stereo TSM methods. Error bars show ±1 standard deviation from the mean.

firmed informal listening tests that the balance of the signal is minimally impacted by stereo TSM. The proposed frame method results in a slightly higher dissimilarity when compared to the whole file method but makes improvements when sound sources move within the stereo field. If the whole file method is used, a sine wave panning from right to left will pan into the centre and back to the right. This was rectified using the frame method at the expense of the computation time for calculating the mid-point and a small increase in dissimilarity for the SPC feature.

Smaller feature extraction frame sizes were tested, and resulted in a minimal increase of dissimilarity for all feature measurements. Synthetic signals were also tested and showed similar, yet less significant improvements.

The mean SPC and balance values for the subset of files at multiple

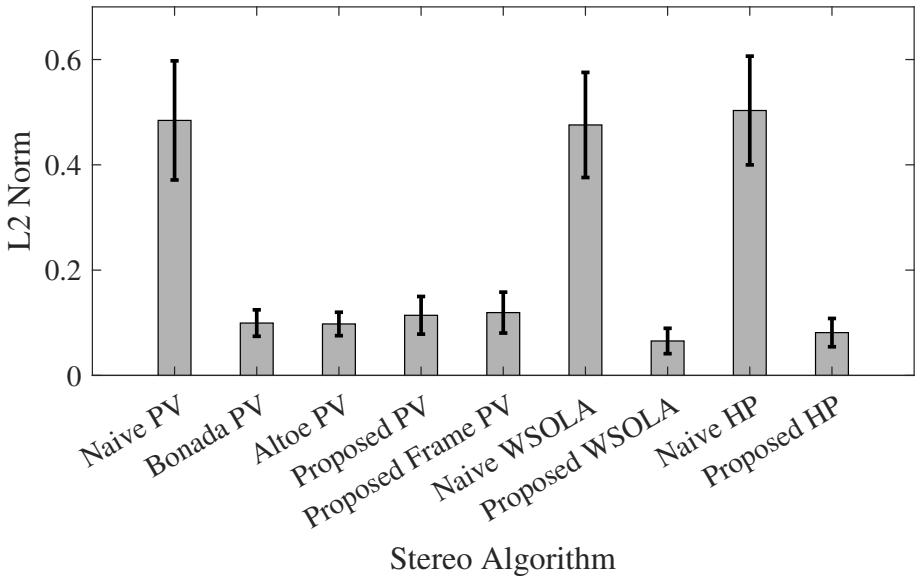

Figure 5.4: Mean SPC dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better.

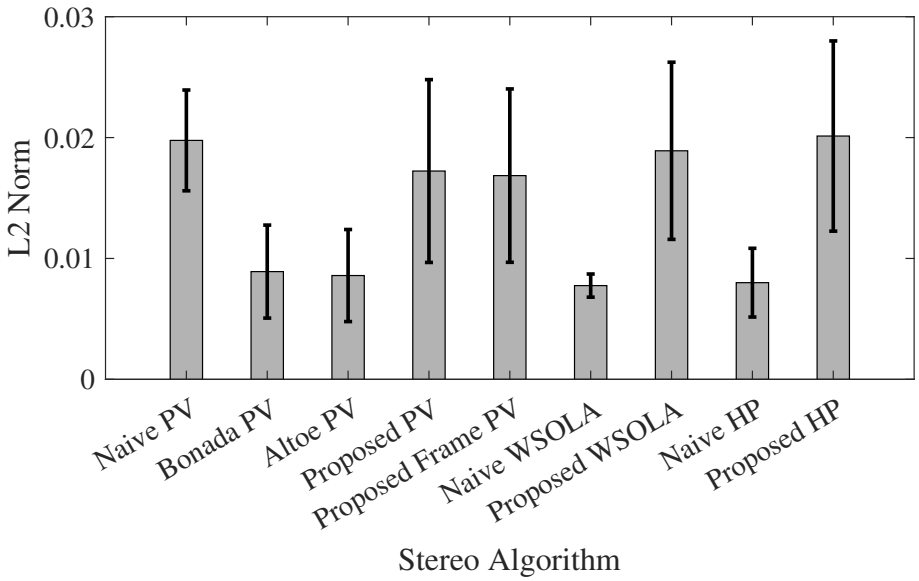

Figure 5.5: Mean stereo balance dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better.

TSM ratios, can be seen in Figures 5.6 and 5.7. Each file is ordered in ascending time-scale ratio, with dotted vertical lines denoting the change in file. The SPC of mono signals are effectively unchanged when doubled across both channels before TSM, with *Male 12* as an example. However, if the recording has been made using multiple microphones, any differences in phase between channels are increased. This is particularly noticeable with the reverberant brass and percussion recording (*Brass and perc 6*). Also of note is the large variability when processing similar sound sources with different recording technique, suggesting that factors other than the source impact on the stereo field, with male speech recordings as a good example. When considering the impact of the TSM ratio, there is a slight inverse relationship for $\beta > 1$ for very coherent source material, however this does not hold true in other cases where there is no discernible trend. An increase in SPC for the Altoe, Bonada and Proposed methods, for source material with a complex sound field, results in a more coherent phase relationship. This gives reasoning to the narrowing of the stereo field.

When considering the balance feature, Figure 5.7, the change after processing is most pronounced when there are a large variety of sound positions in the stereo field, such as *Chiptune*, *Dorothy 2*, *Piano Allegro* and *Yellow 2*. These files respectively contain hard panned sounds, many sound effects filling the sound field, a wide stereo recording of a piano and hard panned violin with flute. Balance of reverberant sounds is improved over the previous class of files with the balance of mono signals maintaining most effectively. It can also be seen that any change in balance tends towards the centre for all methods, with the naive WSOLA method tending further away in rare cases.

The proposed methods show improvement in maintaining the channel

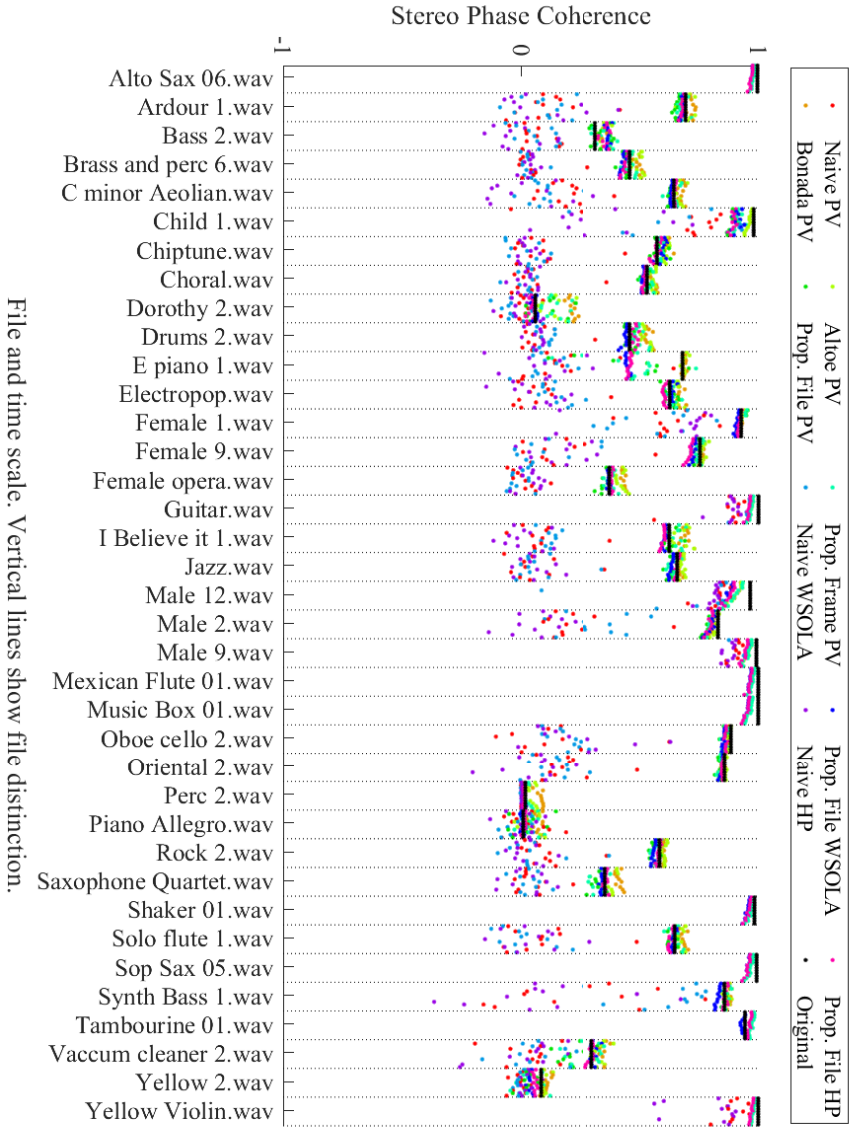

Figure 5.6: [Colour Online] Mean SPC for multiple TSM methods at multiple TSM ratios. Dotted lines segment files. Time-scale ratio increases left to right within each segment.

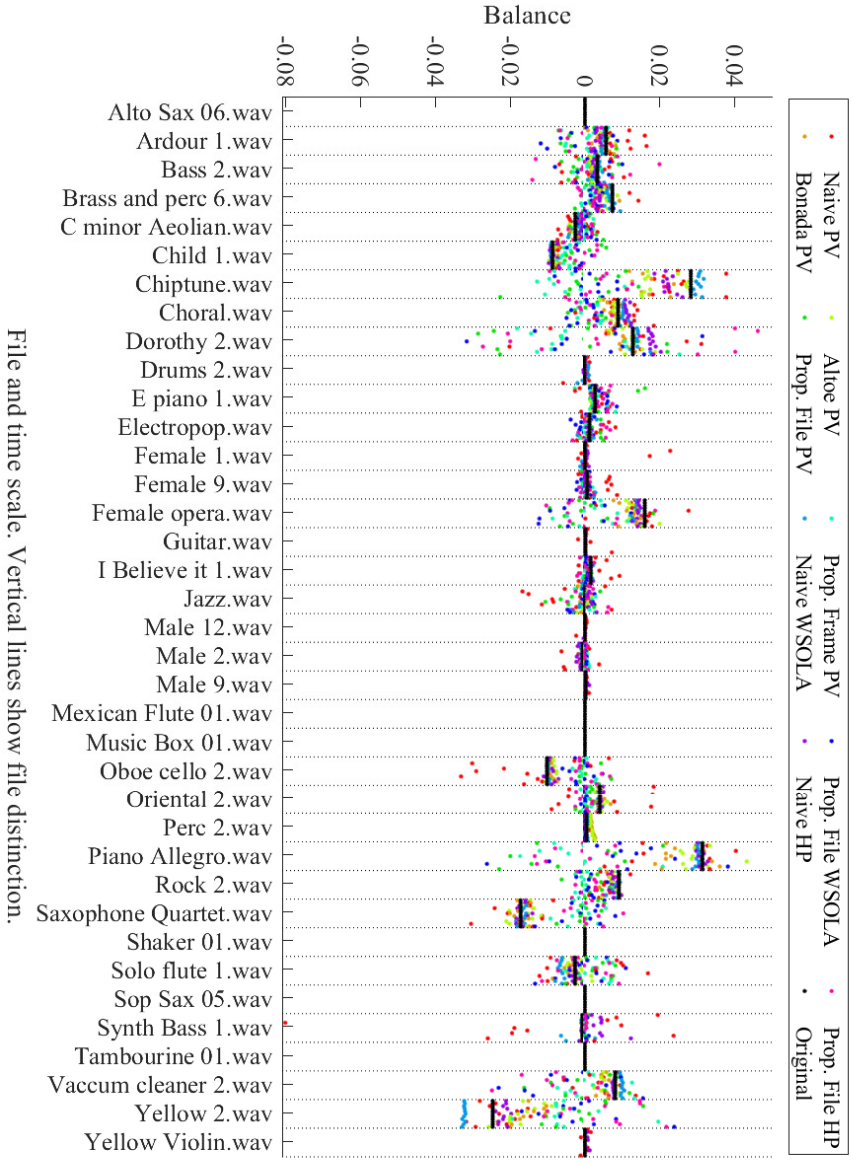

Figure 5.7: [Colour Online] Mean balance for multiple TSM methods at multiple TSM ratios. Dotted lines segment files. Time-scale ratio increases left to right within each segment.

phase relationship, particularly for complex signals and signals containing reverberation. As a result, the presence in the middle of the stereo field is maintained. Due to not explicitly forcing phase relationships to be maintained, and rather causing any drifting out of phase to result in a drift into phase, a narrowing of the stereo field can be heard in some instances. In complex signals such as the Electropop file the central instruments with high spectral energy, such as the bass guitar and kick and snare drums, cause the low spectral energy percussion elements to converge to the centre of the stereo image. While this causes a difference in the width of the signal after time-stretching, the end result is more pleasing to the ear, than the loss of channel phase coherence.

5.1.5 Conclusion

In this chapter two methods of maintaining stereo phase coherence were detailed. These methods used either pre-and post-processing of the entire signal or processing each frame to give real-time suitability. The sum and difference transformation of the stereo signal was calculated before processing and then transformed back after TSM processing. This resulted in a large improvement in stereo phase coherence and maintained the stereo field. The proposed methods produced a high quality stereo output and greatly improved quality over the independent channel processing method, and matched previously published methods. It also allowed for simple implementation around previous TSM implementations, and are suitable for frequency and time-domain TSM methods.

5.2 Fuzzy Epoch-Synchronous Overlap-Add

5.2.1 Introduction

To remove discontinuities and retain phase coherence at the adjusted time-scale, due to adjusting the ratio between the analysis shift size and the synthesis shift size, a number of methods have been proposed. Epoch Synchronous Overlap-Add (ESOLA) is a recent technique that aims to improve the quality of time-scaled speech by extracting glottal pulses, also known as epochs, and using these markers to align the overlap process. By aligning these epoch markers, the primary structure of the signal is preserved in a computationally efficient manner. Additionally, as the epoch locations in the source file are constant, the positions of the epochs need only be generated once, and can then be used for future time-scaling.

5.2.2 Background

The ESOLA method, described in Section 3.2.4 is efficient, however it does not take changes in fundamental frequency into account and is prone to mis-alignment of epochs, shown in Figure 5.8. These errors lead to discontinuities within the time-scaled signal.

The method proposed in this chapter was developed as an additional method for use in the work of Chapter 6. As such, the aim was to create a working implementation with artefacts distinct from previous methods. ESOLA was implemented initially, however results similar to those available online could not be achieved. Occasional loss of pitch information and distortion during changes in fundamental frequency were identified as primary problems to be addressed. As such, the changes

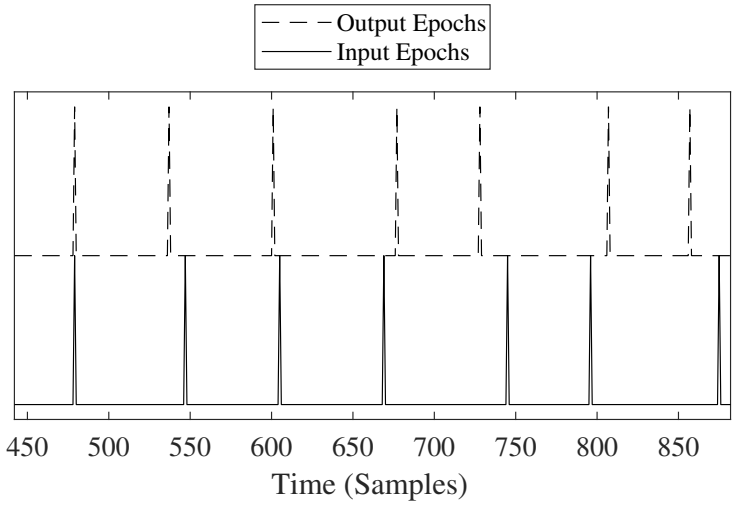

Figure 5.8: Epoch Synchronisation of the ESOLA method.

described below were made to improve the quality of the algorithm.

5.2.3 Method

The proposed method, Fuzzy Epoch-Synchronous Overlap Add (FESOLA), extends the original ESOLA method through the use of cross-correlation when calculating the overlap offset. To enable more effective cross-correlation, the epochs are smeared in the time-domain. Additionally, adaptive window length for epoch extraction was also added, allowing for a greater range of fundamental frequencies. The proposed method is as follows.

An estimate of the fundamental pitch period is found through averaging magnitude spectrum frames across the entire signal and finding the maximum bin location. This bin location is then used to calculate an estimate of the fundamental pitch period. This allows the epoch extraction to be adaptive and suit both male and female voices as well as other non-voice signals. At a sampling frequency of 44.1 kHz, $N_{ZFR} = 217$

was found to be appropriate for male and female speech.

Epochs are extracted using the ZFR algorithm described in Section 3.2.4 and are labelled with a magnitude of 1. Epochs are then made fuzzy through spreading in the time-domain, by setting the samples immediately before and after each epoch to the empirically determined magnitude of 0.6. This process results in a fuzzy pulse train synchronised to the reference signal. Manipulating additional samples around each epoch was explored, however no improvement in optimal frame positioning was found.

To achieve time-scaling, input and output epoch frames are extracted beginning at αS_a and S_s respectively, in the same manner as Rudresh et al. (2018). Cross-correlation between these frames is calculated using

$$R_{xy} = \sum_{m=0}^{L-1} x[m]y[m-k] \quad ; \quad \frac{-3S_s}{4} \leq k \leq \frac{3S_s}{4} \tag{5.14}$$

where L is the length of overlap between frames and S_s is the synthesis shift size. The location of the maximum value within the cross-correlation determines the lead or lag to the start of the adjusted input frame. In the case of multiple maximums in the resultant cross-correlation, the lead or lag closest to the centre of the cross-correlation is used. In the case of either frame containing no epochs, a lag of zero is used.

The frame of samples is extracted starting at αS_a adjusted according to the required lead or lag, and windowed using a Hann window before overlap adding to the output signal. Epochs of the adjusted frame are combined with an output epoch signal using overlap-adding for use in aligning the following frame. An output window signal is also created to allow for normalisation once processing is complete.

5.2.4 Testing Methodology

Small-scale subjective preference testing and large-scale subjective quality testing was undertaken. Small-scale subjective testing compared ESOLA and FESOLA at two different time-scales (β of 0.5358 and 0.7821) for 10 source files containing speech from five female and five male speakers. Testing was undertaken using the Web Audio Evaluation Toolkit (WAET) of Jillings et al. (2015) in an AB format. 10 participants were involved with the testing, all with backgrounds in signal processing.

Large-scale subjective testing was undertaken as part of the work presented in Chapter 6. Participants were presented with the reference and test signals processed using the proposed method, PV, IPL, WSOLA, HPTSM and uTVS, and asked to rate the quality of the processing. 88 source files were scaled at 10 time-scale ratios (β of 0.3838, 0.4427, 0.5383, 0.6524, 0.7821, 0.8258, 0.9961, 1.381, 1.667 and 1.924) resulting in 5280 files. Testing used the WAET with six pairs of files presented per page using horizontal sliders. The number of files in each session was refined during testing, and settled at 60 files per session for a testing time of between 10 and 20 minutes. Approximately 60% of participants were expert listeners, with an average age of 34 and standard deviation of 11 years.

5.2.5 Results

The improvement in epoch alignment for the proposed approach can be seen in Figure 5.9, in comparison to ESOLA in Figure 5.8. The examples are taken from the first input frame containing epochs, of a flute recording, for both methods using the same parameters. This shows that this method is useful is situations beyond time-scaling of speech.

Modifications to the length of ZFR analysis frames must be made if low frequency content is to be time-scaled, to ensure that at least one pitch period or epoch falls within half a frame. The use of a 23ms frame window, in this implementation, results in a half frame pitch period of 11.61ms or 86 Hz.

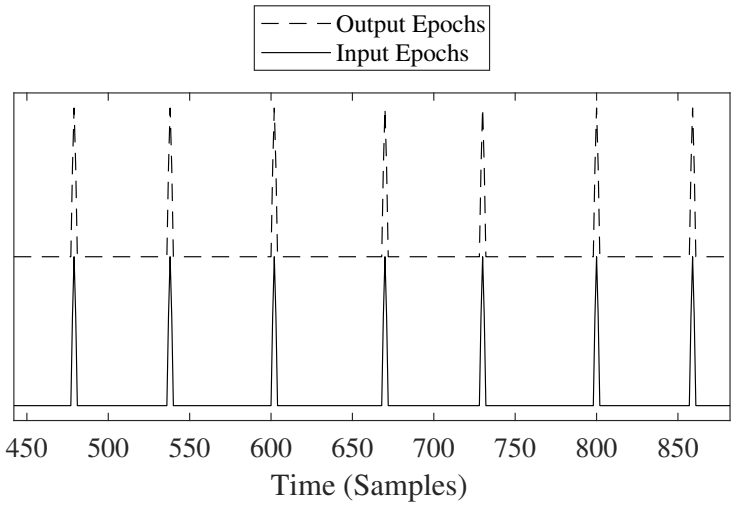

Figure 5.9: Epoch Synchronisation of the proposed FESOLA method.

While viewing the changes between ESOLA and FESOLA using spectrograms can be difficult certain sections can be cropped to allow for closer inspection. Figure 5.10 below shows a short excerpt from the *Child 1* file used in subjective testing. Within this portion is significant pitch inflection, resulting in distortion for the ESOLA method, while there is not distortion when using FESOLA. This is a particularly egregious example, with smaller distortions not immediately visible within spectrograms.

The small-scale preference testing showed a clear preference (77%) towards the proposed algorithm, shown in Figure 5.11. Signals in which the speaker greatly varies their pitch show a stronger preference for the proposed method, while source files with less variation show a more

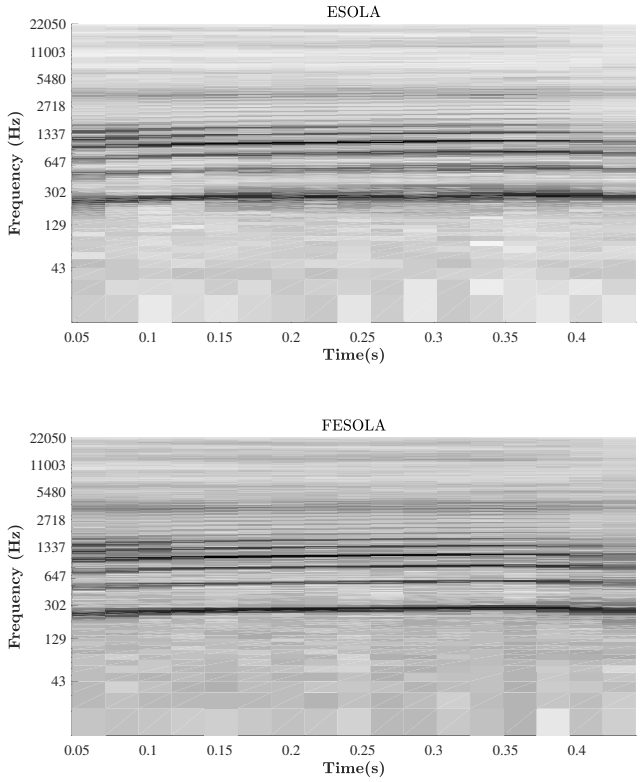

Figure 5.10: ESOLA and FESOLA spectrogram for the *Child 1* file slowed to 53.83%. Distortion around the fundamental and partials are visible for the ESOLA method. Excerpt begins at 4.5 seconds through each file.

even preference between methods. This improvement is consistent with the changes made to the ESOLA algorithm that account for changes in pitch of the speaker, removing distortion for small changes, and reducing distortion for large fast changes.

The large-scale subjective testing results presented in this section are a selection of findings from Chapter 6, containing approximately 19000 signal ratings. The proposed method performs comparatively well for voice signals, Figure 5.12 and solo instrument signals, Figure 5.13.

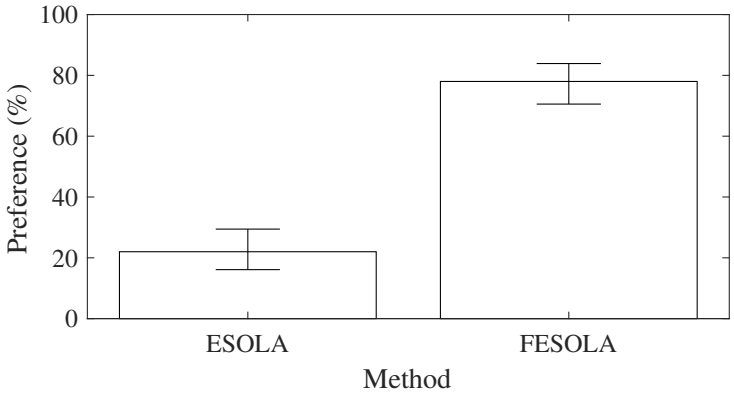

Figure 5.11: Mean preference comparison for ESOLA and FESOLA.

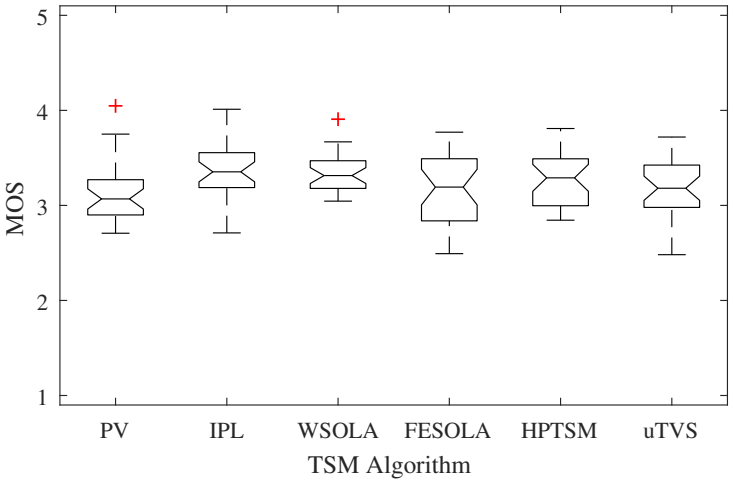

Figure 5.12: Comparison of mean opinion scores averaged across all voice processed files. MOS of 1-5 is Bad-Excellent.

However it gives poor time-scaling for complex musical source material, shown in Figure 5.14. This is due to the reliance on a strong fundamental frequency to allow for generation and alignment of epochs.

Of interest is the relatively poor performance of all methods tested when comparing time-scaling of voice and musical material. This could be due to how often cadence changes within normal conversation. As talking faster or slower is part of general speech, perception may be

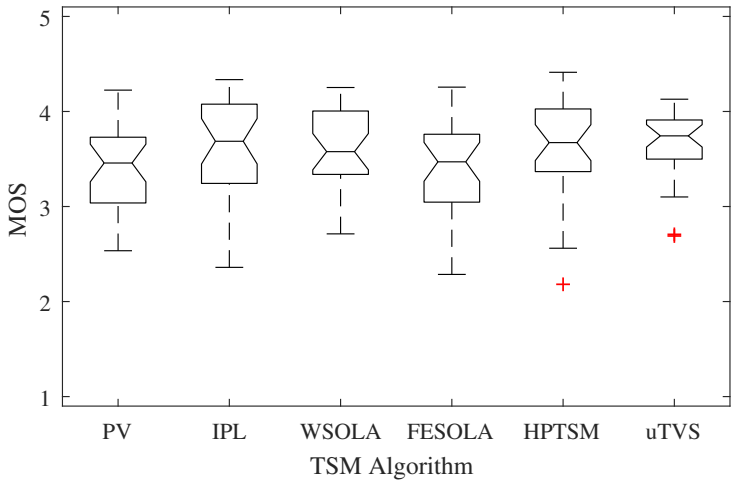

Figure 5.13: Comparison of mean opinion scores averaged across all solo processed files. MOS of 1-5 is Bad-Excellent.

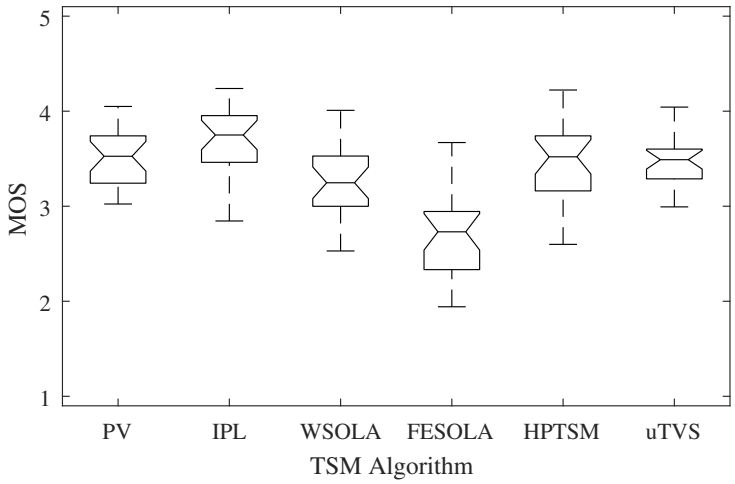

Figure 5.14: Comparison of mean opinion scores averaged across all music processed files. MOS of 1-5 is Bad-Excellent.

more finely tuned for this type of modification. This may then result in a stronger reaction when artefacts corrupt speech signals.

When considering a large variety of source material, including music, solo instruments, sound effects and voice, the quality of the pro-

posed method drops sharply as the time-scale ratio moves away from 100%, shown in Figure 5.15. The falloff is accentuated by poor performance with harmonically complex signals, and while not limited to the proposed method, the drop in quality is more severe than the other methods tested.

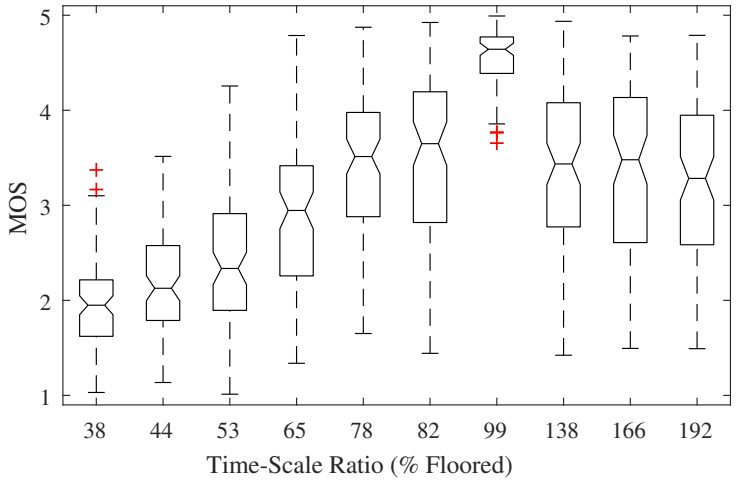

Figure 5.15: Box plots of Mean Opinion Scores for FESOLA across all source files. MOS of 1-5 is Bad-Excellent.

However, the proposed method has similar levels of falloff for voice files when compared to alternative methods. Figure 5.16 shows the mean MOS values for tested methods when processing voice signals across a range of time scales. As will be discussed in Chapter 6, an error was found in the uTVS implementation causing distortion for $\beta \approx 1$.

5.2.6 Conclusion

In this chapter a modification to a TSM algorithm has been proposed. It extends the previous ESOLA method through the use of cross-correlation to align epochs when overlapping frames, and subsequently reduces distortion and artefacts. This change also reduces artefacts due to the

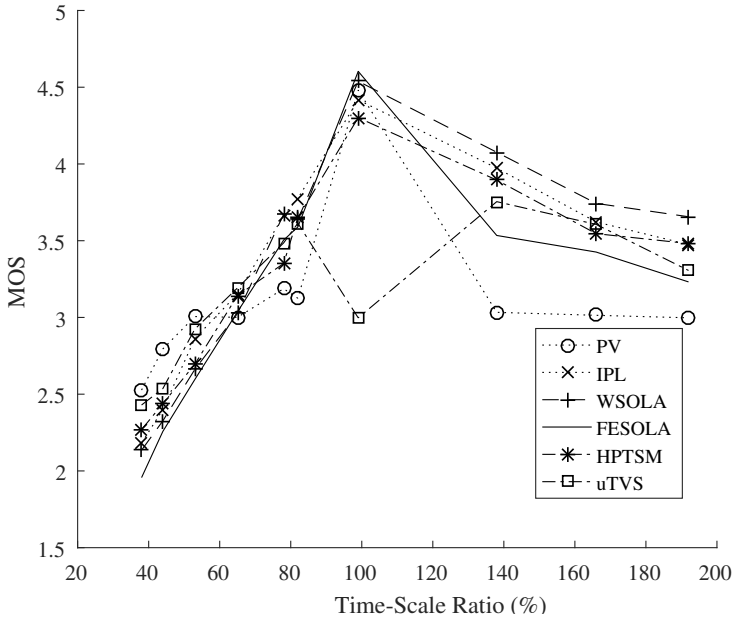

Figure 5.16: Mean MOS for all Voice signals. MOS of 1-5 is Bad-Excellent.

speaker modifying their pitch. The proposed method has been tested against well known TSM algorithms and is found to be preferred over ESOLA, and give similar performance to other TSM algorithms. It was also shown that this algorithm can work effectively with solo instrument signals with strong fundamental frequencies. However, TSM of speech remains an open-problem particularly at slower time-scales.

Chapter 6

A Time-Scale Modification Dataset with Subjective Quality Labels

6.1 Introduction

Despite TSM being a well-researched field, an effective objective measure of quality has not yet been published, limiting comparisons between TSM algorithms. When subjective evaluation has been used, each paper has used a unique set of source material and methods, further reducing comparison to the methods involved in the evaluation. In order to develop an effective objective measure, a dataset with subjective quality labels is required. This chapter details the creation, subjective evaluation and analysis of the first dataset for this purpose, and gives preliminary results for a neural-network-based objective measure of quality.

TSM algorithms most commonly modify the temporal domain by

varying the ratio between analysis (S_a) and synthesis (S_s) shift sizes within an Analysis Modification Synthesis framework. This ratio, given by

$$\beta = \frac{1}{\alpha} = \frac{S_a}{S_s} \tag{6.1}$$

shows α to be the change in signal duration (Roucos and Wilgus, 1985), while β is the playback speed (Sylvestre and Kabal, 1992).

Algorithms for TSM can be classified into three main categories: frequency domain, time domain and hybrid methods. In general, frequency-domain methods excel in scaling harmonically complex material but struggle to produce high quality results with highly transient signals. Time-domain methods are more effective at scaling transient signals but give poor results for polyphonic signals. Hybrid methods leverage the strengths of frequency and time domain methods to produce higher quality results (Driedger et al., 2014).

Common artefacts produced during TSM include 'phasiness' and reverberation (Portnoff, 1981; Laroche and Dolson, 1997), musical and metallic noise or undesirable roughness (Laroche and Dolson, 1999), a buzzy quality (Laroche, 2002) and transient smearing (Laroche and Dolson, 1999). Phasiness and reverberation are heard as a loss of spectral definition and are most commonly associated with frequency domain methods. Laroche and Dolson (1999) suggest that this is due to a change in relationship between the phases of bins in the spectral domain. Musical noise, also known as musical artefacts or musical tones, is due to isolated holes and/or peaks within the power spectrum (Torcoli, 2019). Within TSM, these artefacts are caused by periodicity introduced to noise bins during phase progression, due to the sum of sines model of the Short Time Fourier Transform (STFT). Depending on the frequency relationships between these periodic signals the noise will be perceived

as musical for simple harmonic relationships and metallic for complex harmonic relationships. Transient smearing occurs due to the trade off between STFT spectral and temporal resolution in frequency domain algorithms. As the frame size increases to improve spectral resolution, temporal resolution decreases leading to smearing of transients in time. The buzzy quality, also known as transient skipping or duplication, is an artefact of time-domain methods in which transients may be skipped for $\beta > 1$ or duplicated for $\beta < 1$.

The aim of TSM is often noted, however an exploration of ideal TSM has not been published. For the purpose of subjective evaluation, we describe ideal TSM as indistinguishable from a change by the sound source, that is: the processing should be transparent. A musician changing tempo or a speaker changing cadence would therefore be ideal and should be the goal for TSM algorithms. Consequently, ideal TSM should be determined by the sound source being scaled. For example, a dry recording of individual clicks simply requires temporal realignment of each click, however a recording of sustained notes played on a violin would require the extension of the sustained section of the note's envelope. Further, in the case of a piano, one must consider whether the transient or harmonic nature of the source should be maintained. If a staccato melody played in the upper register without damping is to be slowed, should note decay be lengthened or should the decay be maintained with each note shifted to the new time-scale? We argue that as the piano is a percussive instrument and unable to modify its amplitude envelope, the note decay should be maintained. This is counter to the processing applied by almost all published TSM algorithms. We propose that an ideal TSM algorithm would be sensitive to the signal source and be capable of modifying only the sustained portion of the amplitude

envelope. This raises many questions in the processing of reverberation, vibrato, specific phonemes and more. We consider that content aware or source sensitive TSM is an area with considerable potential for improving the quality of TSM.

The remainder of the chapter is laid out as follows. Section 6.2 briefly describes the TSM algorithms used to create the dataset and previous methodologies for quality evaluation. Section 6.3 describes the source files used in the creation of the dataset and the processing of the source material to create the processed dataset. Section 6.4 describes the subjective testing methodology, opinion score normalisation, results and analysis of the subjective testing and dataset availability. Section 6.5 compares subjective results with published objective measures and provides preliminary results for a novel objective measure of quality. Finally, Section 6.6 summarises and draws conclusions from this research.

6.2 Algorithms and Quality Evaluation

The Phase Vocoder (PV), is a frequency-domain method that uses the known phase progression between frames at the original time-scale to calculate the phase progression between frames at the adjusted time-scale. The digital implementation by Portnoff (1976) uses the STFT to calculate phase spectra and forms the basis for all PV methods published since. The PV is effective at scaling signals with a complex harmonic structure, however it introduces 'phasiness' for non-integer values of α and is prone to transient smearing. See sections 3.3.1 and 3.3.2 for detailed explanation.

The Identity Phase Locking Phase Vocoder (IPL) (Laroche and Dolson, 1999) reduces 'phasiness' introduced by the PV algorithm. The

PV maintains horizontal phase coherence within each STFT bin, however the vertical phase coherence between bins is not maintained. In IPL, the phase of magnitude spectrum peaks are modified, with nearby bins locked to the phase progression of the closest peak. This method was extended, through multi-resolution peak-picking and accounting for added or removed peaks by Karrer et al. (2006). These methods reduce phasiness, however they can introduce a spectral roughness known as metallic or musical noise.

The Waveform Similarity Overlap Add algorithm (WSOLA) (Verhelst and Roelands, 1993) is a time-domain method that uses the similarity between a frame and its natural progression in the input signal to minimize discontinuities in the time-scaled signal. This is in contrast to previous methods that compare with the output signal (Roucos and Wilgus, 1985; Moulines and Charpentier, 1990). WSOLA processes speech and monophonic musical signals effectively, however due to the reliance on the fundamental frequency for alignment, produces low quality results for polyphonic signals.

Fuzzy Epoch Synchronous Overlap-Add (FESOLA) (Roberts and Paliwal, 2019) uses cross-correlation of glottal closure instants, known as epochs, for aligning frames of speech. Epochs are calculated using a Zero Frequency Resonator before smearing in the time-domain. The smearing improves the cross-correlation of epochs, and accounts for changes in fundamental frequency. This method works well for speech and monophonic signals, however it is not effective at processing polyphonic signals.

Harmonic-Percussive Separation Time-Scale Modification (HPTSM) of Driedger et al. (2014) is a hybrid method that uses median filtering of spectrograms for signal separation. WSOLA and IPL are used for percussive and harmonic components respectively. Improved quality was

shown over both individual methods. The method was also shown to compete with contemporary commercial state-of-the-art algorithms.

Multi-component Time-Varying Sinusoidal decomposition (uTVS) (Sharma et al., 2017) uses a Mel-scale filter-bank and the Hilbert transform to calculate instantaneous phase and frequency, bypassing phase unwrapping and the quasi-stationary assumption of traditional frequency-domain methods. As a result, temporal smearing and 'phasiness' artefacts are reduced. This method slightly improves quality over HPTSM, with large improvements over traditional methods.

Elastique (Zplane Development) is a widely used commercial TSM method. While the algorithm is not publicly available, it is currently a state-of-the-art method and has been used in recent TSM subjective evaluations.

Fuzzy classification of spectral bins (FuzzyPV) (Damskägg and Välimäki, 2017), is an extension of the IPL. Spectral bins are given a degree of membership to three classes, sinusoidal, noise and transient, resulting in a fuzzy classification of each bin. Sinusoidal bins are scaled using IPL with phase locking applied to sinusoidal bins, while random phase is added to noise bins. Analysis phases of transients bins are simply relocated in time. Subjective evaluation in Damskägg and Välimäki (2017) shows improvement over HPTSM and similar performance to Elastique.

Non-Negative Matrix Factorization Time-Scale Modification (NMFTSM) by Roma et al. (2019) decomposes the signal into percussive events and harmonic components. Percussive events are copied directly to the output signal, while IPL is used for harmonic components. The duration of percussive events is preserved, however it is highly reliant on correct

detection of the events and introduces novel artefacts.

Little formal subjective testing has been used to evaluate proposed methods, with most proposed methods providing results from informal testing. A wide variety of time-scales and algorithms are used, with little consistency. Time-scales are often limited with two to five times scales ($0.5 \leq \beta \leq 2$) reported in formal testing, with a bias towards $\beta < 1$. This reduces the number of files that require rating, but also limits algorithm evaluation. The difference in quality between $\beta < 1$ and $\beta > 1$ was mentioned briefly by Sylvestre and Kabal (1992). Since the release of the MATLAB TSM Toolbox (Driedger and Muller, 2014), PV, IPL, WSOLA and HPTSM, have been used in most evaluations, while comparisons to commercial algorithms are rare (Karrer et al., 2006; Driedger et al., 2014; Damskägg and Välimäki, 2017). The source audio used during testing also varies between papers with some papers using the files provided with the MATLAB TSM Toolbox. It was noted by Moulines and Laroche (1995) that a thorough perceptual evaluation of TSM approaches had not yet been undertaken.

Two objective measures have been proposed, Signal to Error Ratio (SER) by Roucos and Wilgus (1985) and synthesis consistency (D_M) by (Laroche and Dolson, 1999). SER accounts only for successive magnitude spectra, with no attention paid to phase spectra. D_M also compares the output frame's magnitude to the reconstructed signal's magnitude, however the "measure is not a clear indicator of phasiness" (Laroche and Dolson, 1999). Neither of these measures have seen continued use.

6.3 Dataset Description

The source material for the dataset was collated from my previous creative projects including films, concert and field recordings as well as music written specifically for the dataset. Files were selected to give a broad spectrum of content with variation in TSM difficulty. The number of source files, methods and time-scales was determined by balancing the amount of content required to train a neural network and the number of ratings required for a 'true' Mean Opinion Score (MOS). All content was converted to mono by summing each pair of samples to remove the influence of poor handling of multi-channel files (Roberts and Paliwal, 2018) and normalised to ± 1 before TSM. All files are 16-bit with a sample rate of 44.1kHz and range in SPL from 56.62dB to 86.92dB with a mean and standard deviation of 73.37dB and 6.75dB.

The full dataset contains 34 musical, 37 solo instrument and 37 voice files. Full listings of training and testing subsets are provided in Appendix A and Appendix B, respectively. An additional listing of files extracted but not used can be found in Appendix C The total playback length of the source files is 6 minutes and 42 seconds. Duration was kept short, with a mean of 3.7 seconds and standard deviation of 1.6 seconds, to limit the duration after time-scaling. Files were recorded using a combination of close microphone placement, multi-microphone concert recording, digital synthesis and sampling techniques and shotgun, lapel and large diaphragm condenser microphones. These variations in source material allow for extended subjective evaluation of future TSM methods. The musical and solo files contain synthetic and organic sound sources across classical, rock, jazz, and electronic genres. Voice files contain singing and male, female, and child speech. Finally, the evaluation source files contain a mix of each file type and were used in the genera-

tion of the test and evaluation subsets. Table 6.1 shows an overview of the signal sources.

Table 6.1: Signal sources in each dataset class. All sources within a file are counted separately.

Sound Source	Music	Solo	Voice	Eval
Brass	6	-	-	1
Percussion	7	11	-	2
Piano	6	3	-	2
Rhythm Section	8	4	-	3
Sound Effects	2	1	-	1
String	3	1	-	1
Synthesizers	9	3	-	2
Woodwinds	12	11	-	9
Child	-	-	3	1
Female	-	-	12	3
Male	1	-	15	3
Singing	2	-	4	-
Total per class	27	31	30	20

To form the training set, the source dataset was processed using the first six methods previously mentioned at 10 time-scale ratios resulting in 5,280 processed files. Time-scale ratios (β) of 0.3838, 0.4427, 0.5383, 0.6524, 0.7821, 0.8258, 0.9961, 1.381, 1.667, and 1.924 were generated randomly, but adjusted to ensure coverage across the range of interest. The testing set used Elastique, FuzzyPV and NMFTSM at four random time scales in four bands across $0.25 \leq \beta \leq 2$, resulting in 240 testing files. Subjective evaluation was conducted for both the training and testing sets. An additional evaluation set was created and is discussed in Section 6.5. Full dataset generation took approximately three days on a medium to high end workstation.

The MATLAB TSM Toolbox (Driedger and Muller, 2014) was used with default settings for WSOLA, HPTSM and Elastique time-scaling. FuzzyPV and NMFTSM used provided implementations with default

settings. Author implementations of PV, IPL, uTVS and FESOLA were used with Hann windowing throughout and parameters chosen to maximise informal subjective evaluation. All files were normalised after processing. The PV and IPL used a frame length of 2,048 samples (46.4ms) and synthesis hop of 512 samples. FESOLA used a frame length of 1024 samples (23.2ms). WSOLA used a frame length of 1,024 samples (23.2ms) a synthesis hop of 512 samples and a tolerance of 512 samples. HPTSM used identical IPL parameters while WSOLA had a frame size of 256 samples (5.8ms) and a synthesis hop of 64 samples. uTVS was implemented using six times oversampling and a filterbank containing 88 filters to maintain the relationship between the signal sample rate and filterbank length of the original paper. During testing, an error in the uTVS implementation was found that introduced discontinuities within the instantaneous amplitude and phase during processing at $0.9 \leq \beta \leq 1.1$ for some files. However, as the purpose of the subjective testing was to rate multiple files with a variety of artefacts, they were not removed from the dataset. The error was rectified before creation of the evaluation subset.

6.4 Subjective Testing

Subjective testing was undertaken in two phases. Initial testing was conducted internally within the laboratory. Due to the large number of responses needed per file, testing transitioned to an online browser-based test using the Web Audio Evaluation Tool (WAET) (Jillings et al., 2015). Remote testing greatly increased the number of participants in the study. Participants were contacted in person, directly through social media and email, through mailing lists and public posts on websites such as Reddit and Facebook.

Testing followed ITU-R BS.1248-1 (ITU-T, 2019) recommendations for general methods for the subjective assessment of sound quality as close as practicable, resulting in the following testing parameters. Files were presented in reference-processed pairs with no limits placed on the amount of playback before moving to the next file. Checks were included to ensure both files were played at least once. A continuous grading scale was used in conjunction with a quality scale, where Poor-Excellent corresponds to scores of 1-5. Sessions contained a randomised selection of processed files, presented in random order, with participants free to choose the session they would evaluate. The amount of content per session was refined during testing, for a maximum session duration of 20 minutes. Towards the end of testing, the sessions were restricted to files that had limited responses to reduce MOS standard deviation.

Initial testing was undertaken using a bespoke MATLAB GUI, shown in figure 6.1, that presented individual reference-processed pairs, allowed for saving and restoring of sessions, user input of name, sound transducer, and a check that the participant had no known hearing issues. Participants received training before beginning testing, including explanations of the purpose of TSM and common artefacts with audio examples. A small initial test session of 33 files was completed before a random session was assigned. Each session contained 18 minutes of audio, approximately 200 files, randomly selected from the pool of processed audio files. Participants could elect to evaluate additional sessions following a break equal in length to the completed session.

To increase the number of participants the WAET was used, shown in Figure 6.2. A small number of sessions were evaluated containing 100 files before reduction to 60 files based on participant feedback of session duration. Training identical to laboratory testing was available from

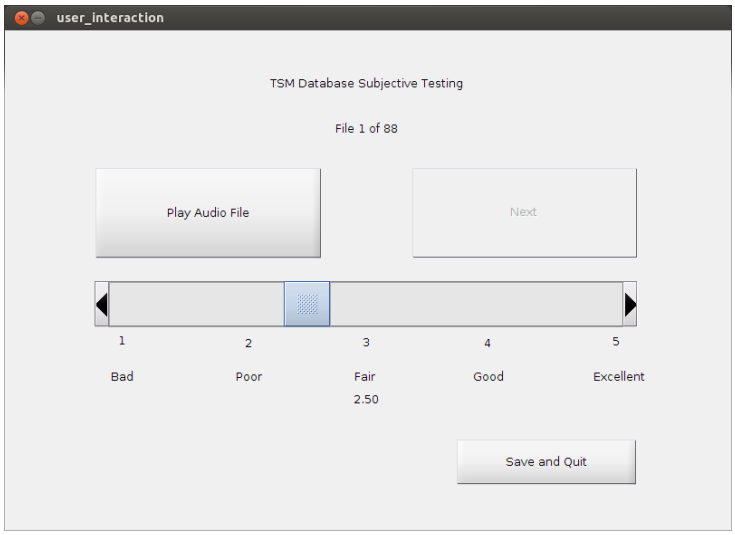

Figure 6.1: Bespoke MATLAB graphical user interface used for laboratory sessions.

the index page, which contained links to each test session. The index page contained reminders to use headphones in a quiet space during testing and a random number generator to suggest which test session the participant should complete. Before each session, name, age, sound transducer, experience in critical evaluation of sound and any known hearing issues were collected. Participants could also elect provide an email address to be contacted for future studies. Each session was split into pages containing six reference-processed pairs.

Figure 6.2: Web Audio Evaluation Tool user interface used for remote testing. Shown with two file pairs.

To remove bias and variability between sessions, opinion scores were normalised according to ITU-R BS1284 (ITU-T, 2019) using

$$Z_i = \frac{x_i - \bar{x}_{si}}{\sigma_{si}}\sigma_s + \bar{x}_s \tag{6.2}$$

where Z_i is the normalised result, x_i is the opinion score of subject i, $\bar{x}_{si}$ is the mean score for subject i in session s, $\bar{x}_s$ is the mean score of all subjects in session s, σ_s is the standard deviation for all subjects in session s and σ_{si} is the standard deviation for subject i in session s.

As the files in each session were unique, means and standard deviations were calculated on the subset of files matching those in the session. Normalised opinion scores were not truncated, however MOS were limited to the subjective interval of 1-5.

6.4.1 Results

A total of 42,529 file ratings were collected from 263 participants across 633 sessions, with 10,354 ratings collected during laboratory testing. Participants ranged in age from 16 to 66 with a median age of 30. 52.36% of ratings were contributed by expert listeners. 12 files were limited to a MOS of 1, while 28 files were limited to a MOS of 5.

Due to the different files and time-scale ratios used for the testing subset, direct comparison between methods in training and testing subsets was not appropriate. However, a general comparison was achieved through local averaging of MOS, centred around training time-scale ratios. Means of adjacent time-scale ratios, bounded by 0.3 and 3, defined the local areas. While 0.3 is greater than some time-scales used within the testing set, it was set empirically to include enough data points, while limiting the impact of much slower time-scales. Mean MOS for

testing subset methods are noisier due to the smaller number of files, and non-uniform difficulty in processing each signal.

Two measures of reliability were used for each session. The Root Mean Squared Error (RMSE) denoted by $\mathcal{L}$ is given by

$$\mathcal{L} = \sqrt{\frac{\sum_{i=1}^{N} (\bar{x}_i - x_i)^2}{N}} \tag{6.3}$$

where the number of files within the session is denoted by N, x_i is the participants opinion score for the file and $\bar{x}_i$ is the overall MOS for the file.

The Pearson Correlation Coefficient (PCC), denoted by ρ, given by

$$\rho = \frac{\mathrm{cov}(\boldsymbol{x}, \bar{\boldsymbol{x}})}{\sigma_{\boldsymbol{x}} \sigma_{\bar{\boldsymbol{x}}}} \tag{6.4}$$

was also used where $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ denote sets of opinion scores and MOS for the session and $\sigma_{\boldsymbol{x}}$ and $\sigma_{\bar{\boldsymbol{x}}}$ are the standard deviation of $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$.

These measures were calculated for each session before and after normalisation. Outliers, calculated prior to normalisation and shown in Figure 6.3, were determined as sessions in which $\mathcal{L}$ or ρ were further than three scaled median absolute deviations away from their respective medians. This resulted in the removal of 45 sessions containing a total of 2,102 ratings (4.94%) from the final pool of sessions.

Following outlier removal and normalisation, $\mathcal{L}$ and ρ means of 0.771 and 0.791 improved to 0.682 and 0.799. Distributions of $\mathcal{L}$ and ρ pre- and post-normalisation can be seen in Figure 6.4.

The use of Intraclass Correlation Coefficients (ICC) (Shrout and Fleiss, 1979; Loizou, 2013) was explored, however as the subjective results are neither fully crossed nor fully nested, ICC cannot be used. In-

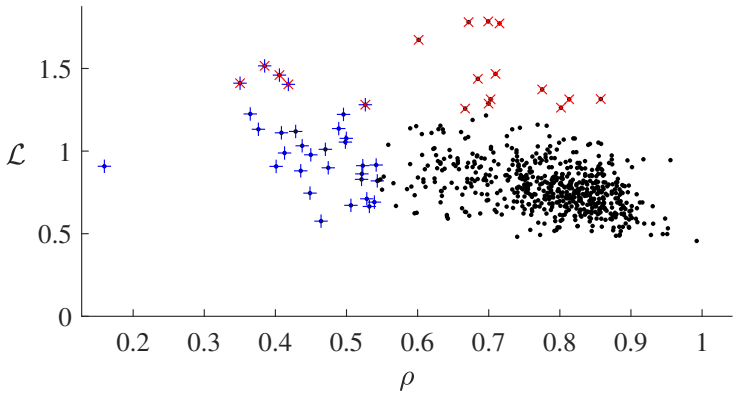

Figure 6.3: [Colour Online] Distribution of PCC and RMSE for all sessions before normalisation and outlier removal. Blue plus symbols mark PCC outliers, while red crosses mark RMSE outliers.

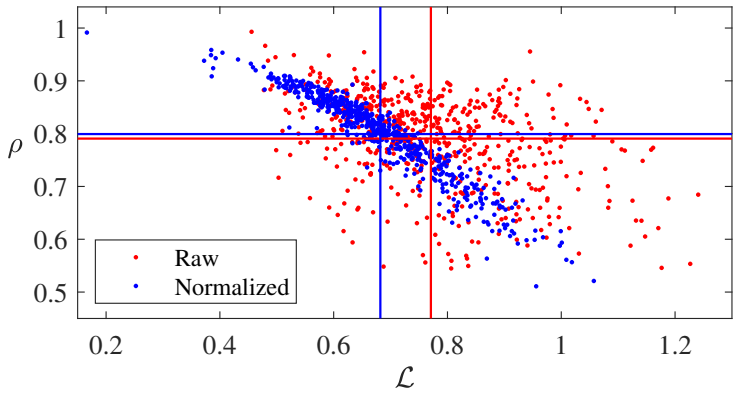

Figure 6.4: [Colour Online] Distribution of PCC and RMSE for each session before normalisation. Horizontal and vertical lines denote means.

stead, the inter-rater reliability for Ill-Structured Measurement Designs of Putka et al. (2008) was used, calculated by

$$G(q,k) = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \left(q\hat{\sigma}_R^2 + \frac{\hat{\sigma}_{TR,e}^2}{\hat{k}}\right)} \tag{6.5}$$

where $\hat{\sigma}_T^2$ is the estimated variance for file main effects (true score), $\hat{\sigma}_R^2$ is the estimated variance for participant main effects, $\hat{\sigma}_{TR,e}^2$ is the estimated variance components for the combination of residual effects and file-participant interaction, and $\hat{k}$ is the harmonic mean of the number of

participants per file. q scales the contribution of $\hat{\sigma}_R^2$ based on the overlap between the sets of participants who rate each file, and is calculated by

$$q = \frac{1}{\hat{k}} - \frac{\sum_i \sum_{i'} \frac{c_{i,i'}}{k_i k_{i'}}}{N_t(N_t - 1)} \tag{6.6}$$

where $c_{i,i'}$ is the number of participants that each pair of files (i, i') share, k_i and $k_i{}'$ are the number of participants who rated files i and i' respectively and N_t is the total number of participants in the sample. This measure gives an overall rater reliability ($G(q, k)$) of 0.871 prior to normalisation and 0.909 post normalisation.

For an overview of all results, Figure 6.5 shows all normalised file ratings ordered by ascending MOS. All opinion scores are shown in the histogram with the overlaid red line showing the MOS for each file. It can be seen that when the TSM quality is very high or very low there is greater consensus amongst participants, however there is a large variance in opinion for files with mid-range quality. It can also be seen that the MOS tracks below the majority of responses in the Good to Excellent range, suggesting a difference between MOS and a majority of opinion scores. Median opinions scores were explored, based on (Jamieson et al., 2004), resulting in tighter groupings, however there was no significant change in averaged scores nor improvement in session reliability. Median opinion scores have nonetheless been included as labels with the dataset, along with mean and median opinion scores calculated before normalisation.

All methods show improvement in quality as β approaches 1, as is to be expected. However, the implementation of uTVS gave poor performance when time-scaling at 0.9961, see Section 6.3, but achieved state-of-the-art performance for all other time-scales. Figure 6.6 shows

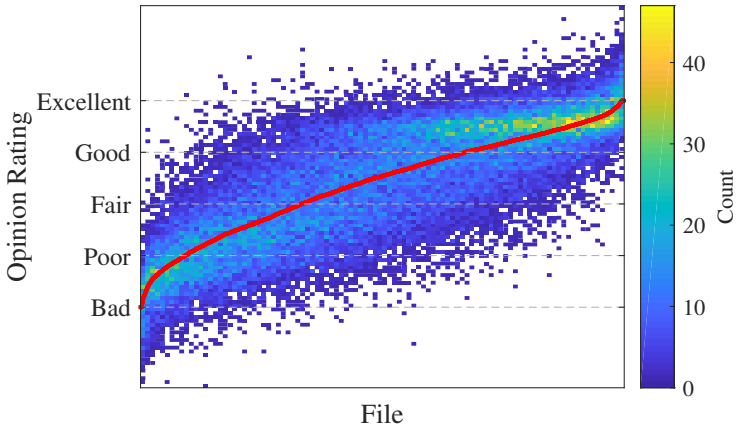

Figure 6.5: [Colour Online] 2D Histogram of normalised responses, ordered by ascending MOS (red line).

the results of each method for each time-scale, averaged across all files. When comparing two inverse time-scale ratios, for example $\beta = 0.5$ and $\beta = 2$, the slower of the pair is lower in quality, suggesting that slowing a file down is perceptually more difficult than increasing its speed. This is consistent with the testing of Sharma et al. (2017), however the effect is more pronounced within this testing. Of interest are two specific cases, that of PV and WSOLA. For $\beta < 1$, PV is perceived to have a higher quality than WSOLA, however this is reversed for $\beta > 1$. It can then be inferred that different artefacts are perceived as having a greater impact on the quality of the TSM. We propose that for $\beta < 1$, the transient-doubling of WSOLA is perceived as worse than the 'phasiness' and transient smearing of the PV, while for $\beta > 1$ transient skipping is less detrimental than the artefacts introduced by the PV. This is a similar finding to Moinet and Dutoit (2011), who noted that some listeners preferred PV artefacts in some cases. Similarly, comparison of PV and IPL shows a change in preference towards the smeary PV artefacts for large reductions in speed, over the metallic artefacts of IPL. The PV was rated comparably to state-of-the-art methods for the

three smallest β.

A surprising result is the high performance of IPL in comparison to HPTSM and uTVS. HPTSM achieved numerically similar results to those given in Driedger et al. (2014). However, while HPTSM was shown to be greater in MOS by 1, our testing found IPL to be rated higher for all except the two slowest time-scale ratios. Artefacts due to harmonic-percussive separation, the use of WSOLA with a very short frame length or the lower sample-rate of the files used in the MATLAB TSM Toolbox may be the cause. Similarly, the reduced sample-rate in original uTVS testing may have contributed to the variance in MOS between testing. Future research should include comparisons between different IPL implementations.

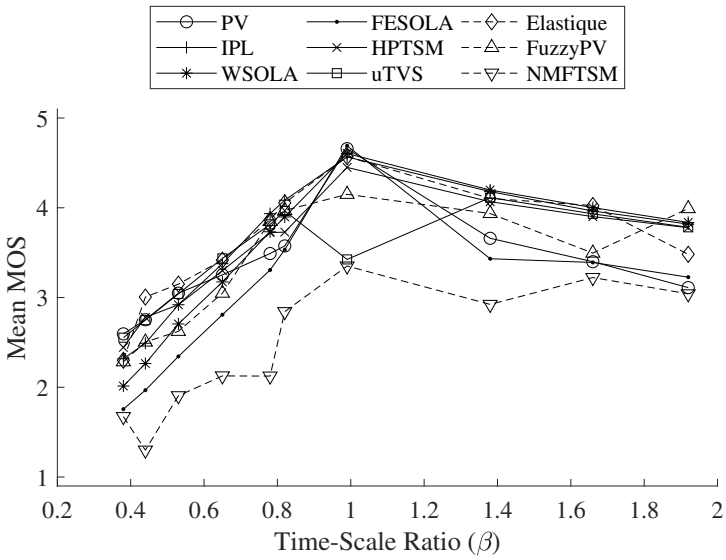

Figure 6.6: Overall means for each method at each time-scale for all evaluated files.

Algorithm performance per class generally follows that of the overall results. As expected however, there are differences in performance quality between methods dependent on the source material. When the mean MOS for each class are considered and $\beta = 0.9961$ results excluded,

Table 6.2: Mean MOS for Overall and Music, Solo Instrument, Voice classes of training source file. MOS for $\beta = 0.9961$ excluded.

	Music	**Solo**	**Voice**	**Overall**
PV	3.450	3.291	2.886	3.202
IPL	3.537	3.636	3.190	3.453
WSOLA	3.133	3.547	**3.262**	3.323
FESOLA	2.418	3.203	2.968	2.882
HPTSM	3.453	3.643	3.118	3.406
uTVS	**3.583**	**3.719**	3.159	**3.486**

uTVS is preferred for music and solo instrument sources while WSOLA is preferred for voice sources. However, the differences in averaged ratings are minor in most cases. Exact mean results are shown in Table 6.2 for TSM methods within the training set.

Perception of processing quality for musical sources, Figure 6.7, confirms the lower quality of time-domain methods, with FESOLA and WSOLA giving poor results. The most interesting result here is that the PV is consistently rated higher than other methods for $\beta < 0.7$ and is comparable for other β. If ratings are averaged for each source file, it is possible to identify 'difficult' files to process. Files with uncorrelated high frequency content were rated poorly, while clean, harmonically simple musical excerpts were rated highly. Signals containing more transient material were rated lower than less transient material. Mean ratings ranged from 2.76 for *Jazz_1.wav* to 3.94 for *Yellow_2.wav*.

Mean MOS results for the solo instrument class of signals, shown in Figure 6.8, improve over musical and voice classes with the exception of the PV for $\beta > 1$. Synthesizer bass sounds were the lowest rated, followed by noisy percussion, polyphonic instruments and tuned percussion, with monophonic harmonic instruments rated highest. The combination of low frequencies with significant transients within the synthesizer bass was particularly troublesome for all TSM methods.

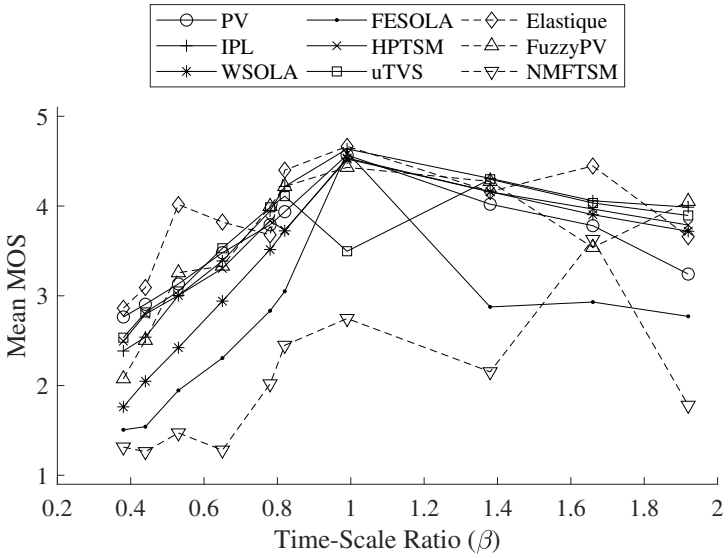

Figure 6.7: Mean MOS for each method at each time-scale for musical source material.

Mean file ratings ranged from 2.54 for *Synth_Bass_1.wav* to 4.17 for *Ocarina_01.wav*.

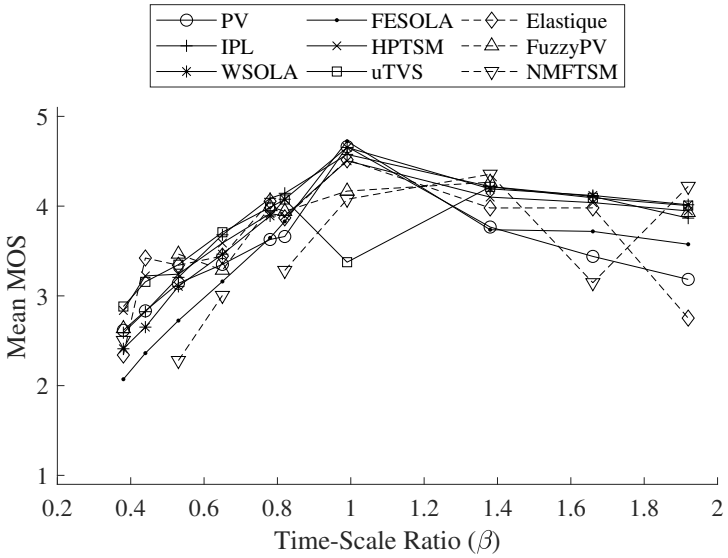

Figure 6.8: Mean MOS for each method at each time-scale for solo instrument source material.

In considering mean MOS for voice signals, shown in Figure 6.9,

WSOLA is preferred for $\beta > 1$, while the preference is less clear for $\beta < 1$. Most methods, except the PV and NMFTSM, were rated similarly for $0.6 < \beta < 1$, however the PV is clearly preferred for $\beta < 0.6$. After this point, smoothness is preferred over transient doubling and metallic artefacts. When considering mean file ratings, the 11 lowest rated files were all male voices, with female and child voices as the seven highest rated files. This mirrors results by Sylvestre and Kabal (1992) who suggested poor frequency resolution for lower frequencies as well as short frame sizes as causes for lower quality. Mean file ratings ranged from 2.73 for *Male_18.wav* to 3.59 for *Child_01.wav*.

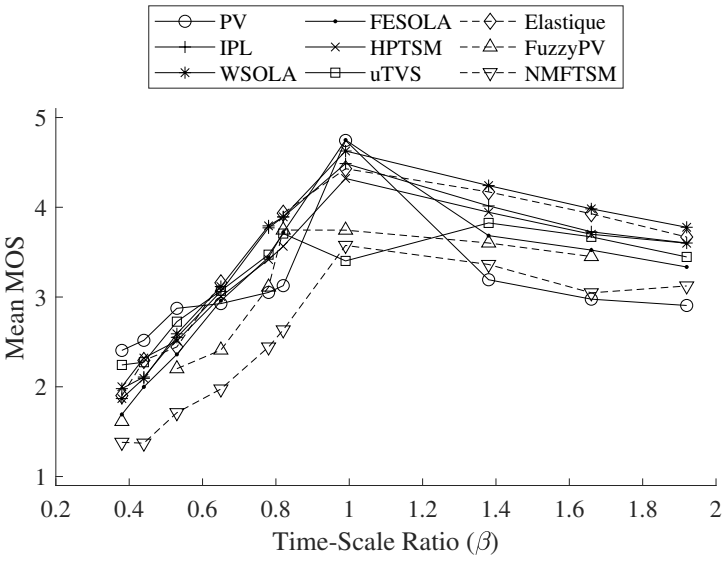

Figure 6.9: Mean MOS for each method at each time-scale for Voice source material.

The mean standard deviation across all files was 0.802 and 0.718, before and after normalisation respectively. As can be seen in Figure 6.10, the range of standard deviation values converges as the number of responses for the file increases. During testing (around 19,000 ratings) this graph showed convergence at around seven ratings per file. As a result, a minimum of seven ratings per file was set as the target to give a 'true'

representation of the quality of the audio file. While there are files that have yet to converge, this is a small subset of the total dataset.

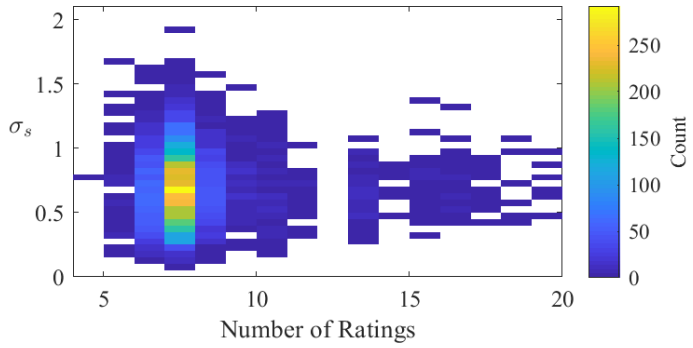

Figure 6.10: [Colour Online] MOS standard deviation against the number of responses for that file.

Comparisons between expert and non-expert listeners, participants with and without known hearing issues and testing modalities were undertaken using the Two One Sided Tests (TOST) of Hauck and Anderson (1984) and Lakens (2017). TOST begins with the null hypothesis of non-equivalent means and uses two one sided tests to show equivalence within a given interval. The interval can be given as a raw score or a standardised difference. If the confidence interval for the difference of the means falls within the equivalence interval, the null hypothesis is rejected and equivalence can be claimed. Analysis was undertaken on session RMSE and PCC values before normalisation, to reduce any interdependence introduced by session normalisation. The equivalence interval was calculated at 5% of the reference sample's mean and Confidence Intervals (CI) of 95% were used throughout. Cohen's sample d is also given for indication of effect size, where d $\approx$ 0.2 is a small effect size.

ITU Recommendation BS.1284 (ITU-T, 2019) recommends investigation of the relationship between expert and non-expert listeners. Par-

ticipants selected if they had experience critically evaluating the quality of audio. RMSE and PCC for non-expert listeners were found to be equivalent to those of expert listeners, with equivalence intervals shown in Figure 6.11. Testing RMSE gave a maximum p-value of 0.0498 and d of 0.1273. Testing PCC gave a maximum p-value of 4.67e-06 and d of 0.1059. We propose that equivalence is a result of the reference-test style of testing and the medium to large impairment in the processed signal, reducing the importance of highly trained critical listening skills for this type of subjective testing.

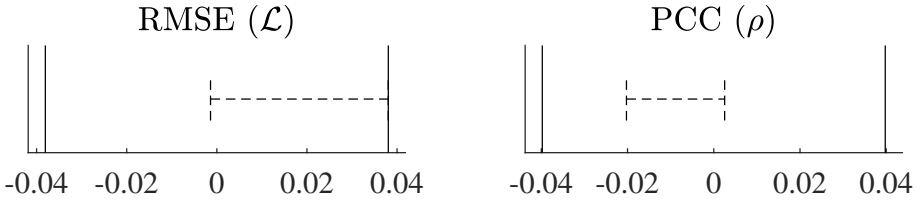

Figure 6.11: TOST (1-α)100% CI for equivalence of participant experience for $\alpha = 0.05$. Equivalence interval of $\pm 5\%$ of expert participant means.

Participants also reported any known hearing issues, with an open answer text box given for responses. Results were not excluded if known issues were reported, but were instead manually sorted into a binary classification of 'No known hearing issues' and 'Any known hearing issues'. Hearing issues included highly descriptive explanations such as "-6dB above 14kHz", a range of tinnitus severity, age related hearing changes and "I like punk music". PCC for participants with any hearing issues were found to be equivalent to those without issue, while RMSE was not found to be equivalent. Equivalence intervals are shown in Figure 6.12. Testing RMSE gave a maximum p-value of 0.2467 and d of 0.0958. Testing PCC gave a maximum p-value of 0.0245 and d of 0.1219. Our proposed explanation is two-fold. Those participants who reported known hearing issues in great detail were also expert listen-

ers, and familiar with the shortcomings of their own auditory system. Additionally, as the participants were presented with the source and processed files and asked to rate the quality of the processing, any issue within the auditory system would affect perception of both files. The small number of sessions classified as 'any issue', 33 compared to 554 for 'no issue', also impacts this result, greatly increasing the standard error. A t-test applied to RMSE was unable to reject that the means are equal with a p-score of 0.4985. Increasing the equivalence interval to ±9.32% allows RMSE equivalence to be claimed. Due to the strong PCC equivalence and close RMSE equivalence, we find no reason to reject sessions in which hearing issues were reported.

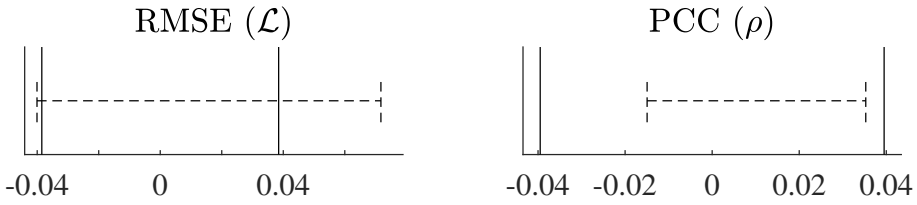

Figure 6.12: TOST (1-α)100% CI for equivalence of means of participants with and without hearing issues for $\alpha = 0.05$. Equivalence interval of ±5% of mean for participants without hearing issues.

As testing was undertaken in different modalities, comparative analysis of results is necessary. PCC for remote participants were found to be equivalent to laboratory participants, while RMSE was not found to be equivalent. Equivalence intervals are shown in Figure 6.13. Testing RMSE gave a maximum p-value of 0.3474 and d of 0.2126. Testing PCC gave a maximum p-value of 0.0013 and d of 0.0931. A t-test applied to RMSE was unable to reject that the means are equal with a p-score of 0.4693. Increasing the equivalence interval to ±8.14% allowed RMSE equivalence to be claimed. Due to the strong PCC equivalence and close RMSE equivalence, we found no reason to reject either testing mode.

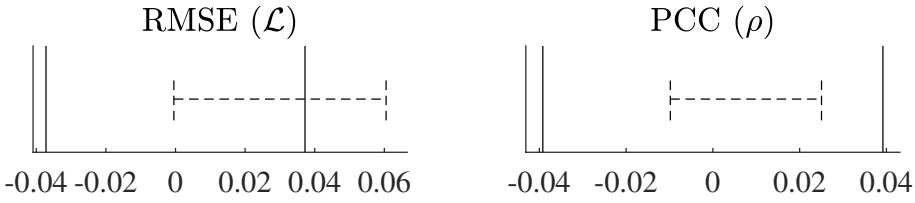

Figure 6.13: TOST (1-α)100% CI for equivalence of testing modality means for $\alpha = 0.05$. Equivalence interval of $\pm 5\%$ of laboratory participant means.

Analysis of the possible impact of age on the quality of the participant's responses was undertaken, the result of which can be seen in figure 6.14. Correlations of 0.108 and -0.001 were found between the age of the participant and the RMSE or PCC respectively, showing no impact of age on evaluation ability.

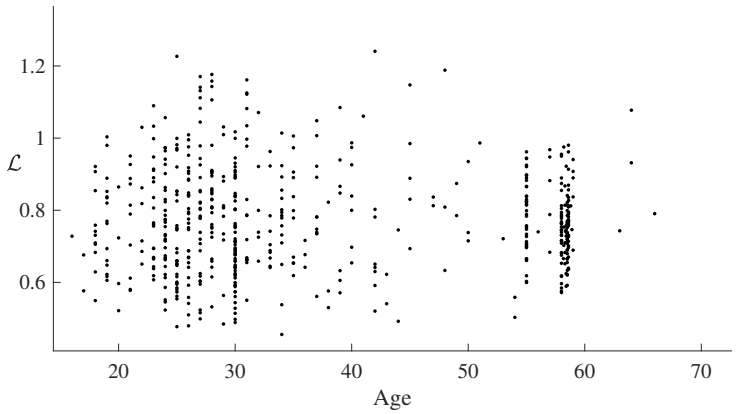

Figure 6.14: Comparison of RMSE ($\mathcal{L}$) and participant age.

The labelled dataset is available, under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, through IEEE-Dataport at http://ieee-dataport.org/1987. Implementation and additional source code is available at github.com/zygurt/TSM.

6.5 Towards an Objective Measure of Quality

Comparison between MOS and previous objective measures, SER and D_M, found correlations of 0.3707 and 0.1574 respectively by averaging absolute correlations for $\beta < 1$ and $\beta > 1$. Signals were aligned through time axis interpolation of the reference magnitude spectrum to the duration of the test spectrum.

Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al., 2000; ITU-T, 2001a) is often used for objective quality evaluation. PEAQ extracts perceptually informed features, using differences between reference and test signals, that are fed into a small neural network to predict subjective scores. Direct application to time-scaled signals is not possible however, due a loss of alignment during TSM. Initial testing, applying the dataset in the design of an objective measure of quality, was undertaken using a modified version of PEAQ. Signals were aligned as above and gave similar correlation to MOS as SER and D_M. The original PEAQ basic neural network was retrained to the subjective MOS, with 10% of the training set reserved for validation. Training used seeds of 0 to 99, with the optimal epoch given by the minimum overall distance ($\mathcal{D}$)

$$\mathcal{D} = \sqrt{\hat{\rho}^2 + \hat{\mathcal{L}}^2} \tag{6.7}$$

where $\hat{\rho}$ and $\hat{\mathcal{L}}$ are calculated by

$$\hat{\rho} = \sqrt{(1 - \overline{\boldsymbol{\rho}})^2 + \Delta\boldsymbol{\rho}^2} \tag{6.8}$$

$$\hat{\mathcal{L}} = \sqrt{\overline{\boldsymbol{\mathcal{L}}}^2 + \Delta\boldsymbol{\mathcal{L}}^2} \tag{6.9}$$

where $\boldsymbol{\rho} = [\rho_{tr}, \rho_{val}, \rho_{te}]$, $\boldsymbol{\mathcal{L}} = [\mathcal{L}_{tr}, \mathcal{L}_{val}, \mathcal{L}_{te}]$, $[.,.]$ denotes concatenation, tr, val and te denote training, validation and testing, $\overline{\boldsymbol{\mathcal{L}}}$ is the mean of $\boldsymbol{\mathcal{L}}$,

$\overline{\boldsymbol{\rho}}$ is the mean of $\boldsymbol{\rho}$, $\Delta\boldsymbol{\rho} = max(\boldsymbol{\rho}) - min(\boldsymbol{\rho})$ and $\Delta\boldsymbol{\mathcal{L}} = max(\boldsymbol{\mathcal{L}}) - min(\boldsymbol{\mathcal{L}})$. The best network achieved a $\mathcal{D}$ of 0.731 and an $\overline{\boldsymbol{\mathcal{L}}}$ of 0.668 and $\overline{\boldsymbol{\rho}}$ of 0.719, placing it at the 11th and 17th percentiles of subjective sessions.

An evaluation set was created by processing the testing subset source files with all methods previously mentioned, at 20 time-scale ratios in the range of $0.22 < \beta < 2.2$. The mean objective output for each method across the range of time-scales is shown in Figure 6.15.

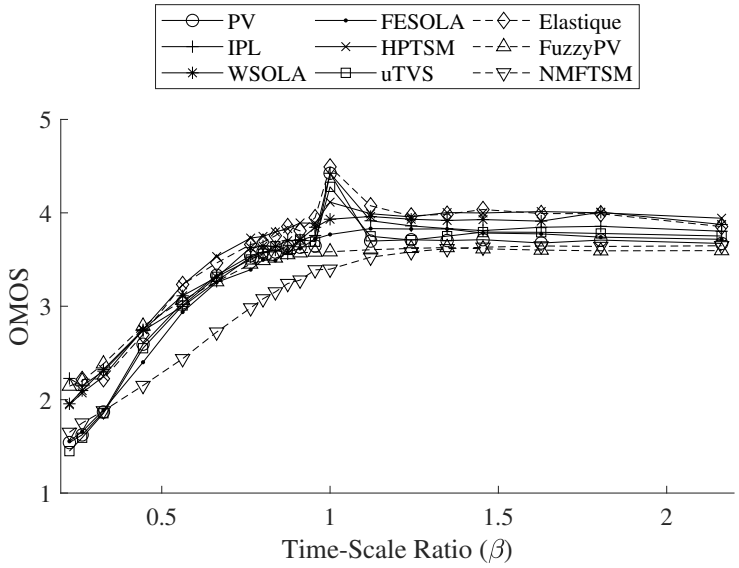

Figure 6.15: Objective MOS for each method in the evaluation set, averaged at each time-scale ratio.

The output exhibits a similar shape to the subjective results, however it only moves away from the mean for $\beta < 0.75$ and $\beta = 1$. Development of an accurate objective measure of quality for TSM algorithms is now achievable, and described in Chapter 7.

6.6 Conclusion

This chapter detailed the creation, subjective evaluation and analysis of a dataset and its use in the development of an objective measure of quality for time-scaled audio. Six TSM methods processed 88 source files at 10 time-scales resulting in 5,280 processed signals for a training subset. Three additional methods at four random time-scales resulted in 240 signals for a testing subset. 42,529 ratings were collected from 633 sessions using laboratory and remote collection methods. Preliminary results for an objective measure of quality were presented, which achieved an RMSE loss of 0.668 and PCC of 0.719. The aim of future work is the design of an improved objective measure of quality for TSM using the dataset, to assist in comparative evaluation of novel methods.

Chapter 7

An Objective Measure of Quality for Time-Scale Modification of Audio

7.1 Introduction

In order to justify the quality of TSM, subjective testing must be undertaken. However, it is expensive and time consuming. Objective methods are available for evaluation of audio quality, however these methods require reference and test signals of identical duration. Consequently, most published objective measures cannot be applied to this context. Two objective measures, SER by Verhelst and Roelands (1993) and D_M by Laroche and Dolson (1999), have been proposed. However, they are shown to be only high level indicators of 'phasiness' or quality (Laroche and Dolson, 1999). In this chapter, we propose the first effective objective measure of quality for time-scale modified audio. It uses handcrafted features with deep-learning methods and is trained using the

dataset of Chapter 6 (Roberts, 2020).

Objective measures of quality seek to predict the quality of a test signal and can be broadly classified into two classes, traditional and machine learning. Traditional measures such as Perceptual Evaluation of Speech Quality of ITU-T (2001b), STOI of Gomez et al. (2011) and the TSM specific measures of SER and D_M are purely analytical in nature. Machine learning methods use neural networks to develop a relationship between subjective evaluations of the test signal and hand-crafted or data driven features extracted from reference and test signals. Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al., 2000; ITU-T, 2001a) is a well known objective measure of this style. Deep learning allows for objective measures that do not require a reference file, as in Avila et al. (2019) for speech quality, however these methods have not yet been applied to TSM.

Training of machine and deep learning methods requires a large amount of labelled signals. For this research we make use of the dataset described in Chapter 6. Reference files were drawn from a large variety of sources including speech, singing, solo harmonic and percussive instruments as well as a variety of musical genres. The training subset, containing 5,280 processed files, was generated using six methods to time-scale 88 reference files at 10 ratios. The methods used were the Phase Vocoder (PV) of Portnoff (1976), the Identity Phase-Locking Phase Vocoder (IPL) of Laroche and Dolson (1999), Waveform Similarity Overlap Add (WSOLA) of Verhelst and Roelands (1993), Fuzzy Epoch Synchronous Overlap-Add (FESOLA) of Roberts and Paliwal (2019), Harmonic Percussive Separation Time-Scale Modification (HPTSM) of Driedger et al. (2014) and Mel-Scale Sub-band Modelling (uTVS) of Sharma et al. (2017). Time-scale ratios ($\beta = \frac{1}{\alpha}$) of 0.3838, 0.4427,

0.5383, 0.6524, 0.7821, 0.8258, 0.9961, 1.381, 1.667, and 1.924 were used for the training subset. The testing subset, containing 240 files, was created using three additional methods to time-scale 20 reference files at a random β in each band of $0.25 < \beta < 0.5$, $0.5 < \beta < 0.8$, $0.8 < \beta < 1$ and $1 < \beta < 2$. Elastique by Zplane Development, the Phase Vocoder using fuzzy classification of bins (FuzzyPV) of Damskägg and Välimäki (2017) and Non-Negative Matrix Factorisation Time-Scale Modification (NMFTSM) of Roma et al. (2019) were used to generate the testing subset. Finally, an evaluation subset was generated by processing the testing subset reference files with all previously mentioned methods, in addition to the Scaled Phase-Locking Phase Vocoder (SPL) of Laroche and Dolson (1999), IPL and SPL variants of PhaVoRIT ($\overline{\text{IPL}}$ and $\overline{\text{SPL}}$) of Karrer et al. (2006) and Epoch Synchronous Overlap-Add (ESOLA) of Rudresh et al. (2018). 20 time-scale ratios in the interval of $0.22 < \beta < 2.2$ were used, resulting in 5,200 files with 400 files per method. During subjective testing 42,529 ratings were collected from 263 participants in 633 sessions resulting in a minimum of 7 ratings per file. Subjective median opinion scores (MedianOS) and subjective mean opinion scores (SMOS) before and after normalisation were provided as labels. The dataset was published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at http://ieee-dataport.org/1987 (Roberts, 2020).

This chapter makes use of the features developed for the basic and advanced versions of Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al., 2000; ITU-T, 2001a). An overall description of the measure is given in Section 4.4.5. This chapter will focus on the modifications made to the generation of features in the context of TSM.

The chapter is organised as follows: Section 7.2 presents the proposed OMOQ method; Section 7.3 presents feature and network results as well

as a comparison of TSM algorithms. Availability, future research and conclusions are presented in Sections 7.4, 7.5 and 7.6 respectively.

7.2 Method

In this section, the proposed TSM objective measure is described. It uses a neural network to infer the SMOS score from hand-crafted features computed from audio processed by TSM. It includes PEAQ features, with modifications described in Section 7.2.1 and additional features specific to TSM artefacts described in Section 7.2.2. Feature preparation is described in Section 7.2.3 and the neural network is described in Section 7.2.4. To give a general overview, a system block diagram can be seen in Figure 7.1. Each group of feature calculations is colour coded, with Reference and Test signals shown as red and blue lines respectively. Black arrows show final feature values, and may denote multiple scalar values. PEAQ feature generation is explained conceptually in Section 4.4.5, with detailed mathematical descriptions in ITU-T (2001a) and Kabal et al. (2002).

7.2.1 Changes to PEAQ

PEAQ was chosen as the starting point for feature generation due to the high level of detail and specificity in the documentation for the measure. Changes were however made to allow for the use of signals of differing lengths, assuming a constant time-scale ratio was applied while processing the signal. Implementation followed ITU-T (2001a), and referred to Kabal et al. (2002) in cases of ambiguity.

Signal preparation begins by summing all input channels and trim-

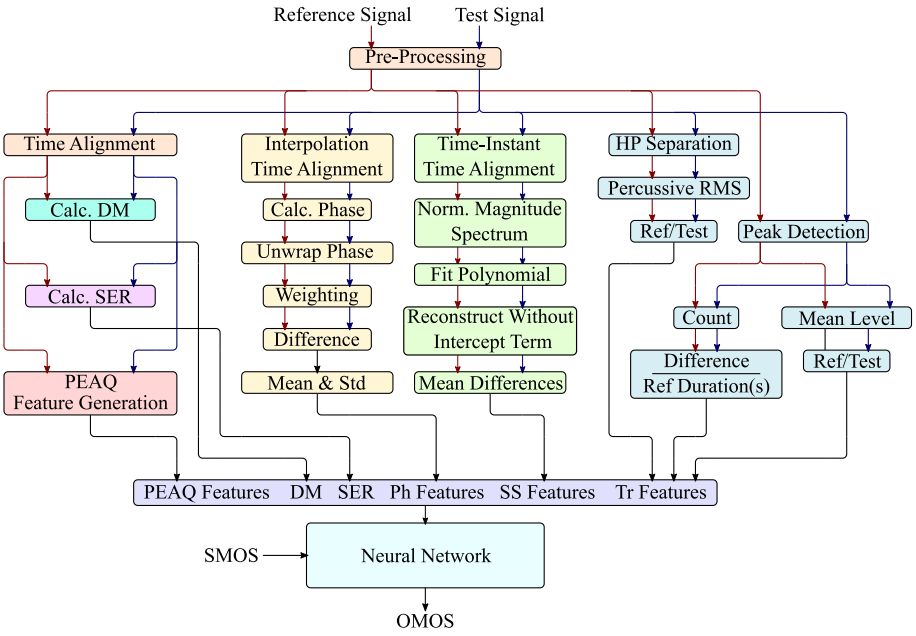

Figure 7.1: [Colour Online] OMOQ system block diagram. Features coloured by group, detail shown for novel features.

ming of silence, before DC removal and normalisation to the maximum absolute value. The proposed method uses full scale as ± 1 rather than the 16-bit integers of PEAQ. A single channel is used in the proposed method as multi-channel TSM is rarely considered (Roberts and Paliwal, 2018). Consequently, a single channel is used for detection probability calculations in ITU-T (2001a) Section 4.7. Silence in the beginning and end of test and reference files is determined as the first and last time the sum of the absolute of four consecutive samples exceeds 0.0061, as per ITU-T (2001a) Section 5.2.4.4. Signals are then truncated to these lengths. This removes frames with low energy at the beginning and end of the signals during averaging calculations, and synchronises the time-scaling starting point.

PEAQ assumes an input sample rate of 48kHz, however the proposed method calculates features based on the sample rate of the input signals.

In the processing of this dataset a sample rate of 44.1kHz is used. As a result, the proposed method uses frequency values calculated from the given bin values of ITU-T (2001a). In the calculation of *BandwidthRefB* and *BandwidthTestB*, frequencies are used rather than bin indices. Noise floor is calculated above 21kHz with 8kHz used as the bandwidth cutoff for bin inclusion during averaging. PEAQ and the proposed method both assume that bandwidth will be reduced due to processing.

The reference signal before and after spectral adaptation, (ITU-T, 2001a) Annex 2 Section 3, is used as input for *AvgLinDistA* calculation. However, the ITU specification is unclear as to which filter envelope modulation ($Mod[k, n]$ in eq. 57 of ITU-T (2001a)) to use in Equation 67. The proposed implementation uses the reference modulation in calculation of s_{ref} and s_{test} for Equation 66 of ITU-T (2001a).

The final change to the ITU standard in the proposed method is the calculation of *RelDistFramesB*. The proposed method uses the interpretation of Kabal et al. (2002) as 'related to' meaning the fraction of frames exceeding 1.5 dB.

Six methods of alignment were investigated during development, time-instance framing anchored to the reference or test signal, and four methods of interpolating magnitude spectrum frequency bins along the time-axis. Time-instance framing extracts frames from the reference and test signals at identical time-instances by scaling the frame locations by β, such that $S_R = u\beta S_T$ where u is the frame number, S_R is the reference signal shift in samples and S_T is the test signal shift in samples. In cases where β is not known, the ratio between the lengths of the truncated input signals is used. A visual representation can be seen in figure 7.2, with frames extracted at the same percentage through the signal.

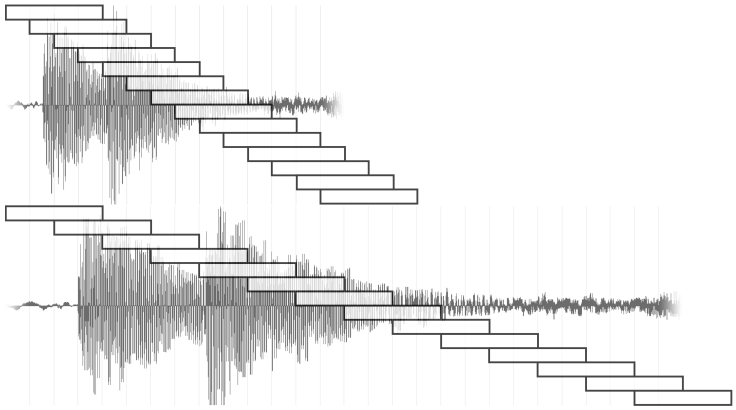

Figure 7.2: Time-instance framing with reference anchor. Reference signal as top waveform and test signal ($\beta = 0.5$) as bottom waveform. Frames start at the same relative position in the signal.

While alignment through re-sampling either the reference or test signal to be the same duration is not suitable, due to resulting changes in pitch, it is possible to re-sample or interpolate low bandwidth representations of the signals, as shown by Sharma et al. (2017). In the proposed method, interpolation of basic PEAQ features is applied prior to the ear model using one of four targets: the longest signal, the shortest signal, the reference signal or the test signal. For advanced PEAQ features, interpolation to the test or reference duration is applied after application of the ear model. There is no requirement for the time-scale to be known during calculation of features, only that the time-scale of the processed signal has been modified by a constant amount. Through a simple thought experiment, we can observe that as we extend signals through interpolation the transient components of the signal will also be extended, while the same transients will not be extended through time-instance framing. As such it is necessary to consider all, and combinations of, alignment methods.

7.2.2 Additional Features

When calculating PEAQ Bandwidth features, asymmetric thresholds are used with +10dB for *BandwidthRefB* and +5dB for *BandwidthTestB*. Test Bandwidth calculated with a +10dB threshold (*BandwidthTest-New*) has been included as an additional feature.

The two published traditional OMOQ were included as features in the proposed method. SER was bound to a maximum of 80 to avoid possible infinite results when processing identical files. This empirical value was the maximum finite feature value for identical files. D_M was used as in Laroche and Dolson (1999).

An equation for the 'phasiness' of non-trivial signals has historically been difficult to develop (Laroche and Dolson, 1999). As such features have been developed to track the resulting differences between reference and test files. One cause of 'phasiness' is phase unwrapping errors that occur when the time-scaling parameter (α) is not an integer (Laroche and Dolson, 1999). In this chapter we propose a method for estimating the level of 'phasiness' by considering the phase progression of reference and test signals. The proposed 'phasiness' features track phase progression through time for reference and test tracks, accounts for the change of time-scale and calculates the difference between the resulting unwrapped phase progression. Weighting is applied to the phase difference, with unity and magnitude spectrum weighting applied in separate features within the proposed method. These features are calculated in the following manner. The phase spectra of the reference and test signals are calculated using the STFT and adjusted to be between 0 and

2π forming $\angle\hat{X}$ using

$$\angle\hat{X}(u,k) = \begin{cases} \angle X(u,k), & \angle X(u,k) > 0 \\ \angle X(u,k) + 2\pi, & otherwise \end{cases} \tag{7.1}$$

forming $\angle\hat{X}(u,k)$. 2π is then successively added to each bin until it is greater than the same frequency bin in the previous frame using

$$\acute{X}(u,k) = \min(\angle\hat{X}(u,k) + 2\pi P) > \angle\hat{X}(u-1,k) \tag{7.2}$$

where $P \in \mathbb{Z}$. The longer $\acute{X}$ is then interpolated to match the length of the shorter signal, forming $\widetilde{X}$. The weighted angle difference ($\Delta\varphi$) can then be calculated using

$$\Delta\varphi(u,k) = \begin{cases} W(k) \cdot (\angle\acute{X}_R(u,k) - \beta\angle\widetilde{X}_T(u,k)), & \mathrm{U}_T >= \mathrm{U}_R \\ W(k) \cdot (\beta\angle\widetilde{X}_R(u,k) - \angle\acute{X}_T(u,k)), & otherwise \end{cases} \tag{7.3}$$

where weighting is calculated with

$$W(k) = \begin{cases} \frac{|X_R(u,k)|}{\max|X_R(u,k)|}, & \mathrm{U}_T >= \mathrm{U}_R \\ \frac{|X_T(u,k)|}{\max|X_T(u,k)|}, & otherwise \end{cases} \tag{7.4}$$

or $W(k) = 1$ for no weighting, where U_R and U_T are the total number of frames in the reference and test signals.

Once the angle differences have been calculated, the mean 'phasiness' features, using No Weighting ($MPhNW$) and Magnitude Weighting ($MPhMW$), are calculated by taking the means of the absolute weighted

difference in time and frequency dimensions,

$$\overline{\Delta\varphi} = \frac{1}{\frac{N}{2}\mathrm{U}} \sum_{k=0}^{\frac{N}{2}} \sum_{u=0}^{\mathrm{U}-1} |\Delta\varphi(u,k)| \tag{7.5}$$

The standard deviation of the absolute weighted difference (*SPhNW* and *SPhMW*) are calculated by taking the standard deviation of the frequency mean of the absolute weighted difference,

$$\sigma_{\Delta\varphi} = \sqrt{\frac{\sum_{i=0}^{\frac{N}{2}} \left(\overline{\Delta\varphi} - \frac{1}{\mathrm{U}} \sum_{u=0}^{\mathrm{U}-1} |\Delta\varphi(u,k)|\right)}{\frac{N}{2} - 1}} \tag{7.6}$$

A number of additional measures were explored including power spectrum weighting, Fletcher-Munson curve weighting and the mean first difference along the time dimension, however they were found to be poor measures or contribute little towards the training of the prediction network.

Figure 7.3 show the 'phasiness' features compared to both SMOS and TSM ratio. 'Phasiness' can be seen to increase as the TSM ratio moves away from 100% and as the SMOS decreases, as expected. Animated 3-dimensional plots rotating between features as functions of SMOS and β, colour coded to each TSM method can be found at zygurt.github.io/TSM/objective.

'Phasiness' causes spectral colouration of the signal (Laroche and Dolson, 1999), allowing for spectral similarity to be used as an indicator of 'phasiness'. Two features (*SSMAD* and *SSMD*) were developed using differences in the smoothed spectrum between reference and test signals. Frames, aligned using reference frame anchors, are converted to normalised magnitude spectra using the STFT and Hann windowing. Third-order polynomials are then fit to the spectra. The resulting

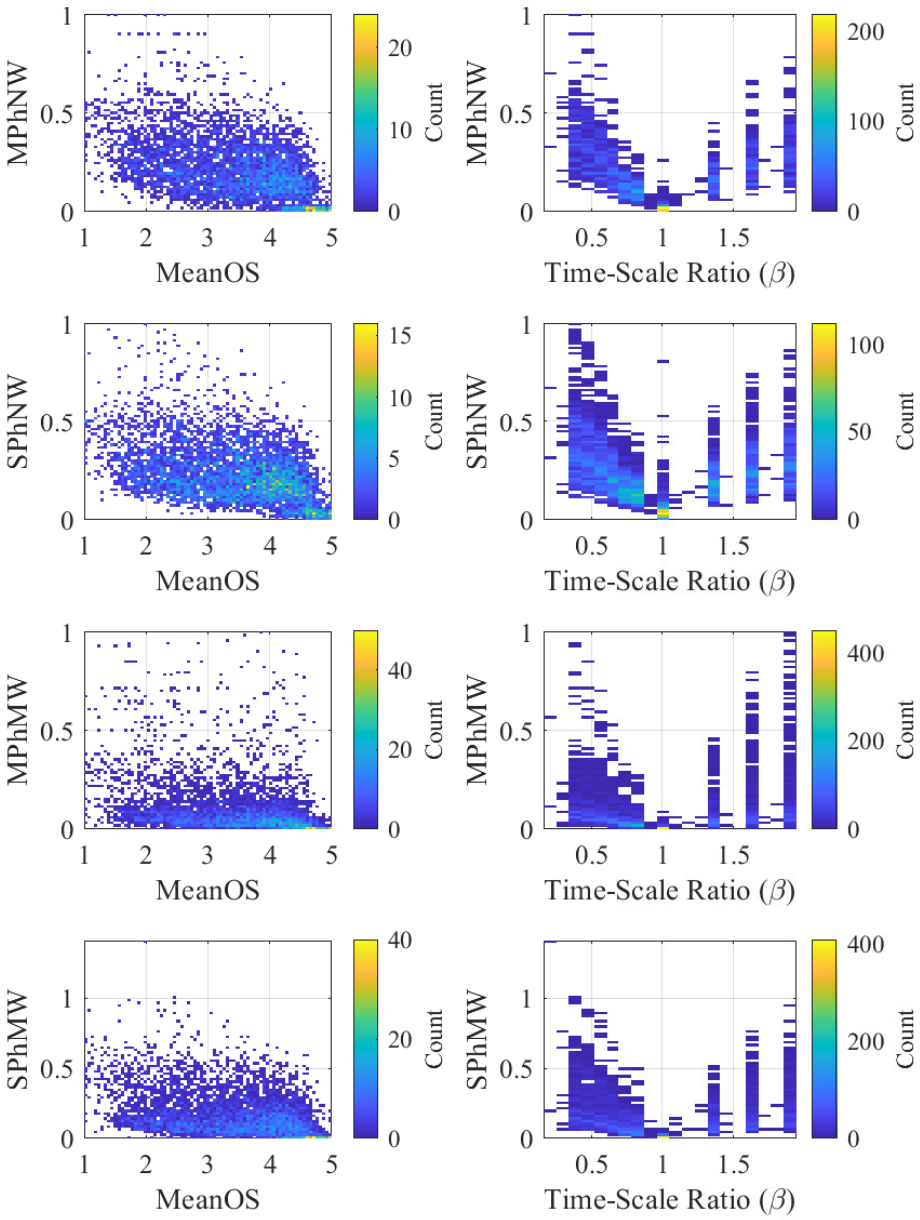

Figure 7.3: [Colour Online] Phasiness features as functions of SMOS and TSM Ratio. Mean/Standard Deviation Phasiness No/Magnitude Weighting ([M/S]Ph[N/M]W).

polynomials, without the intercept term, are applied to a linearly spaced vector $\frac{N}{2}$ in length. Removal of the intercept term removes any overall level difference between the frames. The mean absolute difference and mean difference between reference and test signals are calculated for each frame, with the means of these values forming the two spectral similarity features. These features also give a measure of signal colouration introduced by the TSM algorithm. Figure 7.4 shows the spectral similarity features in relation to the SMOS and TSM ratio. Further analysis found groupings for individual and classes of TSM methods within the features. Time-domain methods inherently introduce less or no 'phasiness', and FESOLA and WSOLA tend to have better spectral similarity than frequency-domain methods. Refer to Figure 7.5 and 7.6 and the *SSMAD* and *SSMD* animated graphs in the supplementary material[1] for examples.

Changes in the transient content of the signal are common TSM artefacts. Three features have been developed for the proposed method, Peak Delta, Transient Ratio and Harmonic Percussive Separation Transient Ratio, with no requirement for alignment between signals. Peak Delta (ΔP) is the difference in the number of onsets between the reference and test signals per second. Onset detection is applied to both signals using the spectral features method described by Bello et al. (2005). A weighting function, $W[k] = |k|$ in

$$\widetilde{E}[u] = \sum_{k=0}^{\frac{N}{2}-1} W[k]|X|^2 \tag{7.7}$$

is applied to the power spectrum before the first backward difference of

[1] See https://zygurt.github.io/TSM/objective for animated graphs of features.

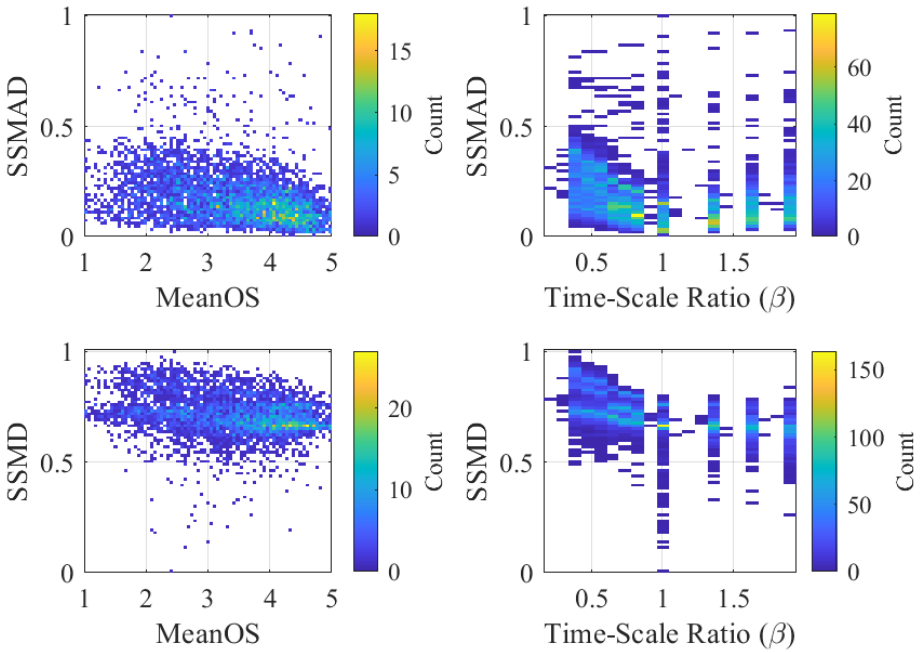

Figure 7.4: [Colour Online] Spectral similarity features as functions of SMOS and TSM Ratio.

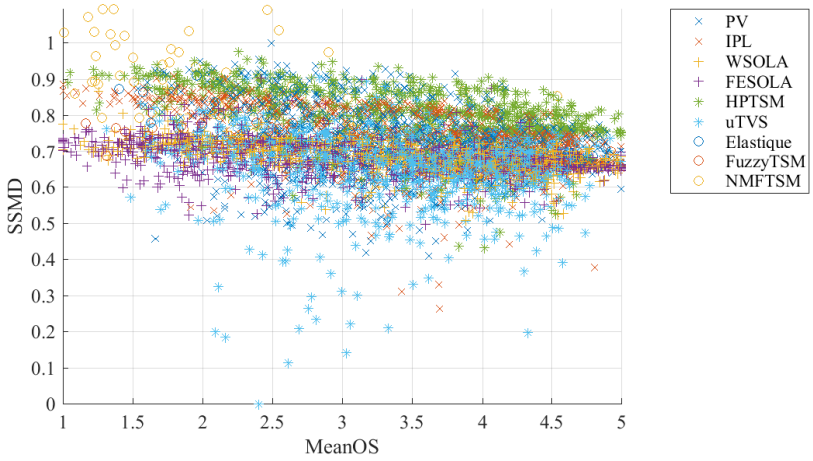

Figure 7.5: [Colour Online] Spectral similarity mean difference feature as functions of SMOS and TSM method.

the logarithmic transform is calculated using

$$\Delta\widetilde{E}[u] = \log_{10}\widetilde{E}[u] - \log_{10}\widetilde{E}[u-1] \tag{7.8}$$

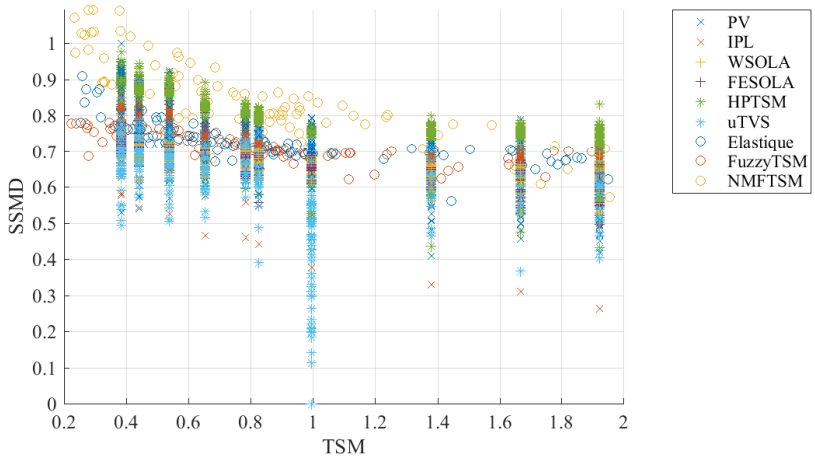

Figure 7.6: [Colour Online] Spectral similarity mean difference feature as functions of TSM Ratio and TSM method.

Peak picking is applied to the onset results, where a peak is greater than its four surrounding values, with

$$\mathrm{P}[u] = \begin{cases} 1, & \Delta\widetilde{E}[u] > \Delta\widetilde{E}[u-2:u+2] \\ 0, & otherwise \end{cases} \tag{7.9}$$

Finally, the difference in the number of peaks per second, calculated using

$$\Delta\mathrm{P} = \frac{f_s}{dim(x_R)} \left(\sum \mathrm{P}_T[u] - \sum \mathrm{P}_R[u] \right) \tag{7.10}$$

is used as the feature, where f_s is the sampling frequency and $dim(x_R)$ is the length of the reference signal in samples.

The transient ratio ($TrRat$) is a measure of the change in transients due to processing and makes use of the peak locations calculated previously in Equation 7.9. It is calculated by selecting peaks where the onset peak level is greater than one standard deviation above the mean

onset level using

$$\hat{\mathrm{P}} = \mathrm{P}, \quad \text{where} \quad \Delta\widetilde{E}[\mathrm{P}] > \overline{\Delta\widetilde{E}} + \sigma_{\Delta\widetilde{E}} \tag{7.11}$$

Peak values are then used to calculate the ratio of mean transient level between the reference and test signals using

$$\mathrm{TrRat} = \frac{\frac{1}{\mathrm{V}_R}\sum_{v=0}^{\mathrm{V}_R-1}\Delta\widetilde{E}_R[\mathrm{P}[v]]}{\frac{1}{\mathrm{V}_T}\sum_{v=0}^{\mathrm{V}_T-1}\Delta\widetilde{E}_T[\mathrm{P}[v]]} \tag{7.12}$$

where V is the total number of selected peaks, and v is the index of the selected peaks.

The Harmonic Percussive Separation Transient Ratio (*HPSTrRat*) compares the Root Mean Square (RMS) levels of reference and test transients. Transients are extracted from reference and test signals using the median filtering method of Driedger et al. (2014). The RMS of the extracted signals are calculated before the final feature is computed by the ratio of reference to test. Figure 7.7 compares each of the transient features to SMOS and TSM ratio.

As musical noise is a known artefact introduced by TSM, it was also explored as a possible feature. Spectral Kurtosis, as proposed by Torcoli (2019), was explored using all previously discussed methods of alignment. Lower, middle and upper frequency bands were used in addition to the maximum across all bands. As all time-alignment methods produced highly correlated results, interpolation to test was chosen as the alignment method. However, inclusion of these features reduced neural network performance and as a result were removed from the features used in the final proposed network.

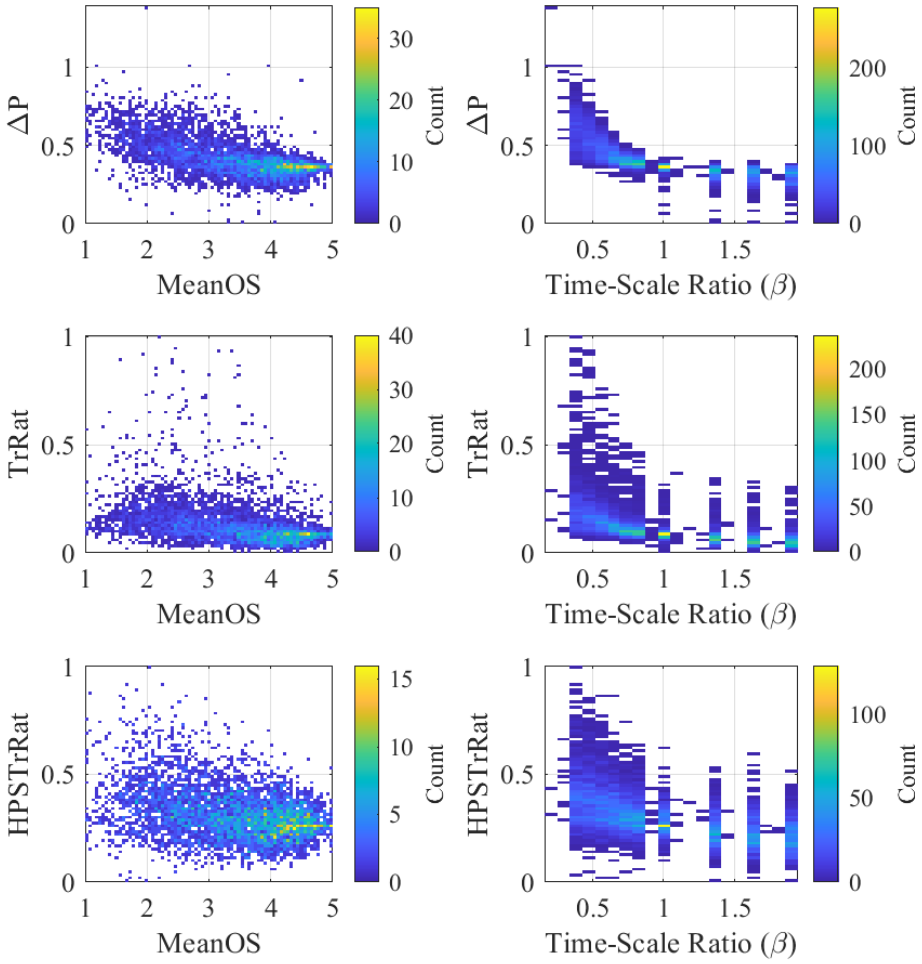

Figure 7.7: [Colour Online] Transient features as functions of SMOS and TSM Ratio.

7.2.3 Feature Preparation

Prior to network training, features were scaled to the interval [0,1] using

$$SMOS \leftarrow \frac{SMOS - 1}{4} \tag{7.13}$$

where a_{min} and a_{max} are the minimum and maximum values for each feature. Target scores were also scaled to the interval [0,1] using equation 7.13. Normalisation of features to zero mean and unit standard

deviation was also explored, however no improvement in network performance was found.

7.2.4 Network Structure

Estimation of opinion scores was formulated as a regression problem using a fully-connected neural network with three hidden layers of 128 nodes, shown in figure 7.8. Layer normalisation and ReLU activation were used with residual connections around the second and third layers facilitated by adding the input of a layer to its output. Sigmoid activation was applied to the final output. The network has 36737 trainable parameters.

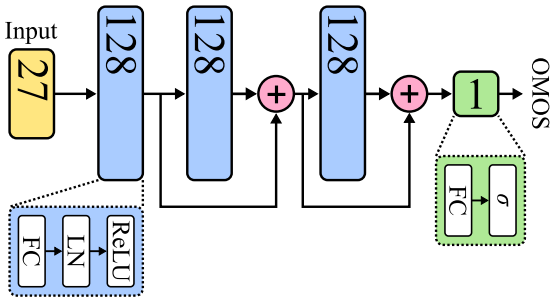

Figure 7.8: Neural network of proposed measure. Numbers denote number of layer output nodes, FC is a Fully Connected layer, LN is Layer Normalization, ReLU activation function, $\oplus$ is element-wise summation of layer input and output values and σ denotes a sigmoid activation layer.

10% of the training dataset was reserved for validation. The network was trained for 800 epochs using a single batch, RMSE loss ($\mathcal{L}$), AdamW optimisation (Loshchilov and Hutter, 2017) and a learning rate of $1e^{-4}$. Networks that were still improving after 800 epochs were trained for an additional 800 epochs. Internal loss values were calculated using estimates in the interval of [0,1], while reported loss values were calculated using estimates scaled back to the original interval of [1,5]. Pearson

correlation coefficient (ρ) and $\mathcal{L}$ were used as network performance measures. As prediction of opinion scores for novel TSM methods is the network aim, early stopping based on validation loss was not used. The optimal epoch was chosen as the epoch with the minimum overall distance ($\mathcal{D}$), calculated using the method described in Section 6.5.

This allowed for the novel artefacts of the testing subset to inform the chosen optimal network, without their use in training the network.

7.3 Results

7.3.1 Feature Results

An initial larger set of features was heuristically optimised based on changes in network performance and similarity to other features, to reduce redundant features. Features were manually pruned if the Pearson Correlation Coefficient (PCC) between features of the same type was above approximately 0.95. This pruning increased the performance of the trained network. Features removed include Spectral Kurtosis features, alternative weightings of 'phasiness' features and most standard deviations of time-domain features before averaging. Due to the non-linear nature of the relationship between β and SMOS, correlation was calculated separately for $\beta < 1$ and $\beta > 1$ before averaging. Figure 7.9 showing the correlation between each of the features in the proposed measure for $\beta < 1$, Figure 7.10 shows $\beta > 1$ and Figure 7.11 shows the average correlation.

Correlation with SMOS for $\beta < 1$ ranges from 0.0024 to 0.6675 with an average correlation of 0.2705, while correlation with the time-scale ratio ranges from 0.0052 to 0.7775 with an average of 0.3128.

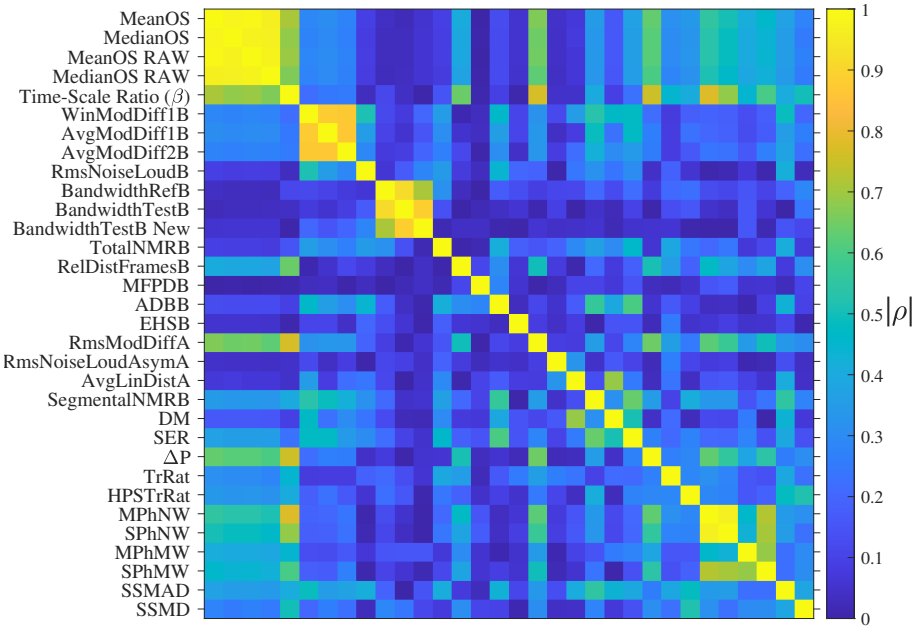

Figure 7.9: [Colour Online] Feature correlation matrix for final features. Absolute correlation for $\beta < 1$.

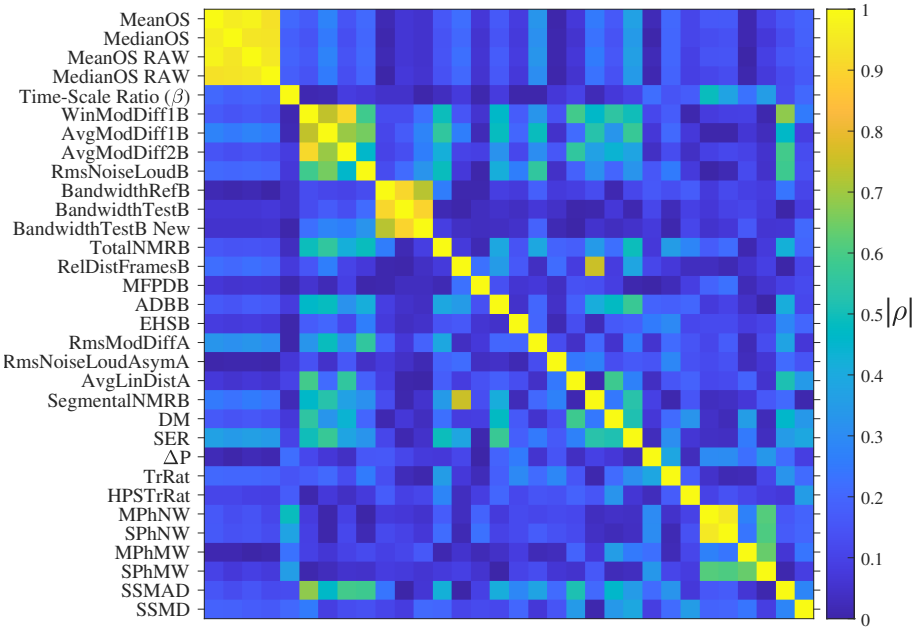

Figure 7.10: [Colour Online] Feature correlation matrix for final features. Absolute correlation for $\beta > 1$.

Correlation to SMOS for $\beta < 1$ ranges from 0.0075 to 0.3705 with an average correlation of 0.1463, while correlation with the time-scale ratio ranges from 0.0046 to 0.4889 with an average of 0.1183.

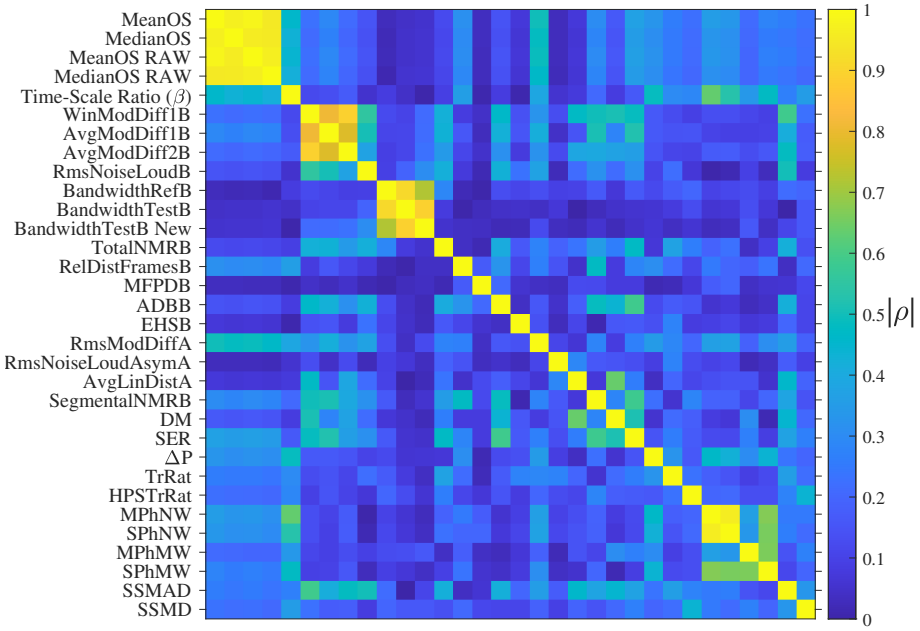

Figure 7.11: [Colour Online] Feature correlation matrix for final features. Average absolute correlation for $\beta < 1$ and $\beta > 1$ shown due to non-monotonic nature of relationship between features and time-scale ratio.

Average correlation to SMOS ranges from 0.0276 to 0.4979 with a mean of 0.2084, while correlation with the time-scale ratio ranges from 0.0056 to 0.6332 with an average of 0.2155. The additional features were found to have a greater correlation to the SMOS and TSM than most PEAQ features. Of interest is the lack of individual features highly correlated with the SMOS or β, while still resulting in excellent network performance. Features were generated at approximately 400 files per hour using 16 threads on a Xeon E5-2630.

7.3.2 Network Performance

A wide range of testing and network configurations were considered during the development of the proposed method. Network hyper-parameters were optimised through a systematic non-exhaustive search. Each method of alignment was trained to SMOS, MedianOS, raw SMOS and raw MedianOS targets, where raw values were calculated prior to session normalisation. Additionally, baseline conditions, the inclusion of reference files within the training set, concatenation of logarithmic transforms of features and combinations of multiple alignment methods were considered. Deterministic training of the network was conducted using seeds from 0 to 99. Figure 7.12 shows the box plot distribution of the best $\mathcal{D}$ for each of the seed values used while training to SMOS. Lower values are better, with a smaller range meaning less reliance on the initial seed.

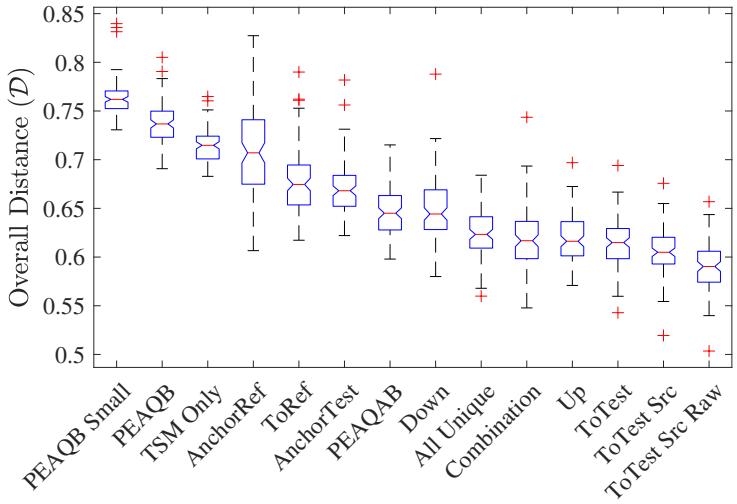

Figure 7.12: Box plot of best distance measure for each seed and training configuration ordered by median $\mathcal{D}$. PEAQB NN uses original PEAQ network, all others use the network described in Section 7.2.4. Lower is better, less spread means less reliance on initial seed.

Across all test cases, the network was more successful training to

Table 7.1: RSME loss mean ($\overline{\mathcal{L}}$) and range ($\Delta\mathcal{L}$), PCC mean ($\overline{\rho}$) and range ($\Delta\rho$), median overall distance ($\widetilde{\mathcal{D}}$) and minimum overall distance ($\min(\mathcal{D})$). Trained to SMOS unless specified. Best results in bold.

Features (Alignment)	$\overline{\mathcal{L}}$	$\Delta\mathcal{L}$	$\overline{\rho}$	$\Delta\rho$	$\widetilde{\mathcal{D}}$	$\min(\mathcal{D})$
Original PEAQB (To Test)	0.668	0.054	0.719	0.075	0.762	0.731
PEAQB (To Test)	0.636	0.104	0.753	0.028	0.737	0.691
TSM Only (To Test)	0.630	0.115	0.764	0.026	0.715	0.683
All (Anchor Ref)	0.540	0.205	0.834	0.086	0.707	0.607
All (To Ref)	0.549	0.203	0.827	0.093	0.675	0.617
All (Anchor Test)	0.524	0.268	0.842	0.124	0.668	0.622
PEAQAB (To Test)	0.558	0.109	0.820	0.043	0.645	0.598
All (To Shorter)	0.543	0.120	0.836	**0.024**	0.644	0.580
All Unique(All Alignments)	0.524	0.117	0.844	0.036	0.623	0.560
Combination (To Test & Anchor Test)	0.477	0.221	**0.873**	0.085	0.617	0.548
All (To Longer)	0.534	0.109	0.834	0.030	0.616	0.571
All (To Test)	0.500	0.150	0.860	0.050	0.615	0.543
All (To Test Incl. Ref)	0.490	0.101	0.864	0.030	0.605	0.519
All (To Test Incl Ref) (SMOS Raw)	**0.474**	**0.089**	0.859	0.028	**0.590**	**0.503**

mean, rather than median, targets. Consequently, the results discussed below will be solely focused on networks trained to mean targets. To increase readability, median overall distance ($\widetilde{\mathcal{D}}$) and the best case $\boldsymbol{\mathcal{D}}$ with associated $\overline{\boldsymbol{\mathcal{L}}}$, $\Delta\boldsymbol{\mathcal{L}}$, $\overline{\boldsymbol{\rho}}$ and $\Delta\boldsymbol{\rho}$ values can be found in Table 7.1. Values were calculated as per Section 7.2.4.

The baseline performance for the traditional methods was determined by correlation with the target. SER and D_M gave overall ρ with subjective scores of 0.3708 and 0.1574 respectively. Machine learning baseline performance was obtained by applying time-aligned PEAQB features to the original PEAQB network described by ITU-T (2001a), shown as 'Original PEAQB (To Test)'. By increasing the complexity of the network, to that in Section 7.2.4, $\overline{\boldsymbol{\mathcal{L}}}$ and $\overline{\boldsymbol{\rho}}$ were improved, shown as 'PEAQB (To Test)'. Performance was further improved through the inclusion of PEAQ Advanced features, shown as 'PEAQAB (To Test)'. Interpolating to the length of the test signal was found to give the best performance followed by, in order, interpolating up to the longer signal, down to the shorter signal, anchoring frame locations to the test signal, interpolating to the reference signal length and anchoring frame locations to the reference signal. Using only the new TSM features gave improved performance over the PEAQ Basic features. Including reference signals as test material, with targets set to 5, improved network performance and gave the best overall distance for the seeds tested. Combinations of features generated using interpolation to test and time-instance anchoring to test (Combination) were also applied to the network. This improved best performance over each alignment individually, with training performance sensitive to changes in a single feature and is highly reliant on initial seed selection. All Unique features were also combined and applied to a larger network with 512 nodes per layer, but did not

improve over previously tested feature sets. Finally, combinations of concatenating logarithmic features, including reference signals and combining different alignment features were applied to the network, but all resulted in reduced performance. The loss and correlation for each epoch of the proposed network can be seen in Fig. 7.13.

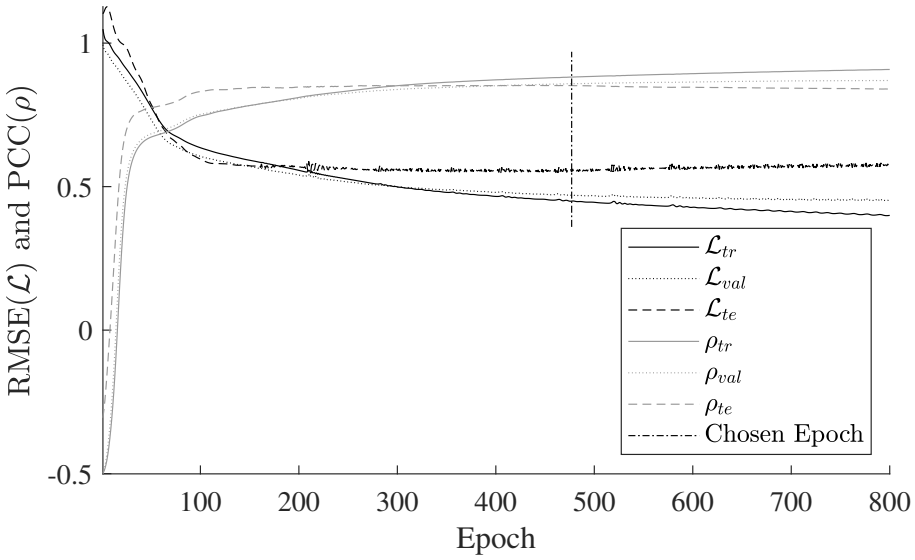

Figure 7.13: Loss and Correlation for training, validation and test sets for each epoch. Best epoch shown as vertical line.

Given the network performance in predicting raw SMOS outperforms prediction of normalised SMOS, investigation of Objective Mean Opinion Score (OMOS) differences was undertaken. The mean difference between normalised and raw SMOS was found to be -0.0023, while the mean difference was found to be 0.016 for OMOS. Normalising was found to slightly extend the range of the SMOS values, with higher ratings for high quality files, and lower ratings for low quality files. Given the ITU-T (2019) recommendation of normalisation, the final proposed objective measure of quality was trained to normalised SMOS using features aligned using interpolation to test, including reference files.

The proposed network achieved a best mean PCC of 0.864 and RMSE of 0.490, training to normalised SMOS using interpolating to test for alignment and including reference files within the training set. These results place the proposed network at the 82nd and 97th percentiles of subjective sessions for PCC and RMSE respectively.

7.3.3 TSM Algorithm Evaluation

TSM algorithms were compared using the evaluation subset, described in Section 7.1. The uTVS implementation used in subjective testing ($\overline{\text{uTVS}}$), and an IPL by Driedger and Muller (2014)(DIPL) have also been included. Although $\beta = 1$ was used in the evaluation, in practice time-scaling is only applied at ratios other than 1. Additionally, $\beta = 0.25$ is the minimum available for Elastique. Consequently, all results for $\beta = 1$ and $\beta < 0.25$ were excluded from averaging calculations. Table 7.2 shows the mean OMOS for each of the TSM methods tested in addition to means for each file class ordered by ascending overall mean.

Analysis is split into each class of reference file followed by overall average results. The poor performance of the uTVS subjective testing implementation, for β close to 1, is also visible, with the updated implementation showing monotonic improvement towards $\beta = 1$. In all cases, the noisy nature of the results for testing TSM method in Chapter 6 has been smoothed.

For musical files, the OMOQ effectively differentiates between frequency and time-domain methods, where the quality worsens faster for time-domain methods. WSOLA fairs the best of time-domain methods, diverging from frequency domain methods for $\beta < 0.8$, as shown in figure 7.14(a). When averaged, the OMOQ rates IPL highest followed by

Table 7.2: Mean OMOS for each class of file and overall result. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$.

TSM Method	Music	Solo	Voice	Overall
NMFTSM	2.931	2.932	2.987	2.948
ESOLA	2.782	3.635	3.302	3.194
FESOLA	2.987	3.652	3.412	3.314
PV	3.572	3.463	3.052	3.383
FuzzyPV	3.663	3.399	3.160	3.433
Phavorit SPL	3.678	3.586	3.201	3.507
uTVS	3.591	3.628	3.320	3.521
Subjective uTVS	3.615	3.630	3.317	3.530
Phavorit IPL	3.715	3.663	3.212	3.548
HPTSM	3.597	3.673	3.415	3.565
SPL	3.672	3.646	3.457	3.600
WSOLA	3.538	3.792	3.507	3.605
Elastique	3.721	3.796	**3.660**	3.725
Driedger's IPL	3.771	**3.850**	3.596	3.742
IPL	**3.835**	3.773	3.621	**3.752**

Elastique. All other frequency domain methods gave similar results.

For solo files all methods except NMFTSM perform similarly with a maximum difference between methods of 0.576 for $\beta = 0.87$. Method means at each time-scale can be seen in figure 7.14(b). Driedger's IPL has the highest mean OMOS, followed by Elastique, WSOLA and IPL as shown in Table 7.2. The strong performance of WSOLA is expected, due to individual harmonic and percussive signals.

Voice file OMOS shows the greatest variance between methods. Of interest is the exponential shape of the curve for $\beta < 1$ compared to the logarithmic shape for musical and solo classes, indicating harsher subjective evaluation of voice files was learned by the network. Method means at each time-scale can be seen in figure 7.14(c). Elastique has the highest mean OMOS, followed by IPL, DIPL and WSOLA. ESOLA and FESOLA give improved performance for this class relative to other

methods.

By averaging all OMOS, IPL has the highest average rating followed by DIPL and Elastique, separated by only 0.03 OMOS. Only 0.098 separates WSOLA through $\overline{SPL}$. The overall low performance of FuzzyPV is unexpected, given that it builds on IPL. However other methods that perform decomposition of the signal, such as NMFTSM and HPTSM, also perform below the methods they build upon, suggesting that simpler artefacts are preferred over those introduced by multiple processing methods. The overall means can be seen in figure 7.14(d). Two-sample t-test analysis ($\alpha = 0.05$) of all OMOS shows the null hypothesis of equal means to be rejected in almost all cases when the absolute difference of mean OMOS is greater than 0.098. ESOLA and FESOLA are the only exception with an absolute difference of 0.1201 and P-value of 0.069.

7.4 Availability

The proposed tool is available from github.com/zygurt/TSM. This includes the MATLAB scripts for feature generation, PyTorch code feature evaluation and features for all dataset files in 'csv' and '.mat' formats. A bash script is also included that creates a virtual environment and installs required modules. The features are also available with the subjective dataset at http://ieee-dataport.org/1987.

7.5 Future Research

Future research is multi-faceted. Evaluation of a wide range of commercial and lesser known published TSM methods should be considered in addition to comparisons of different implementations of the same TSM

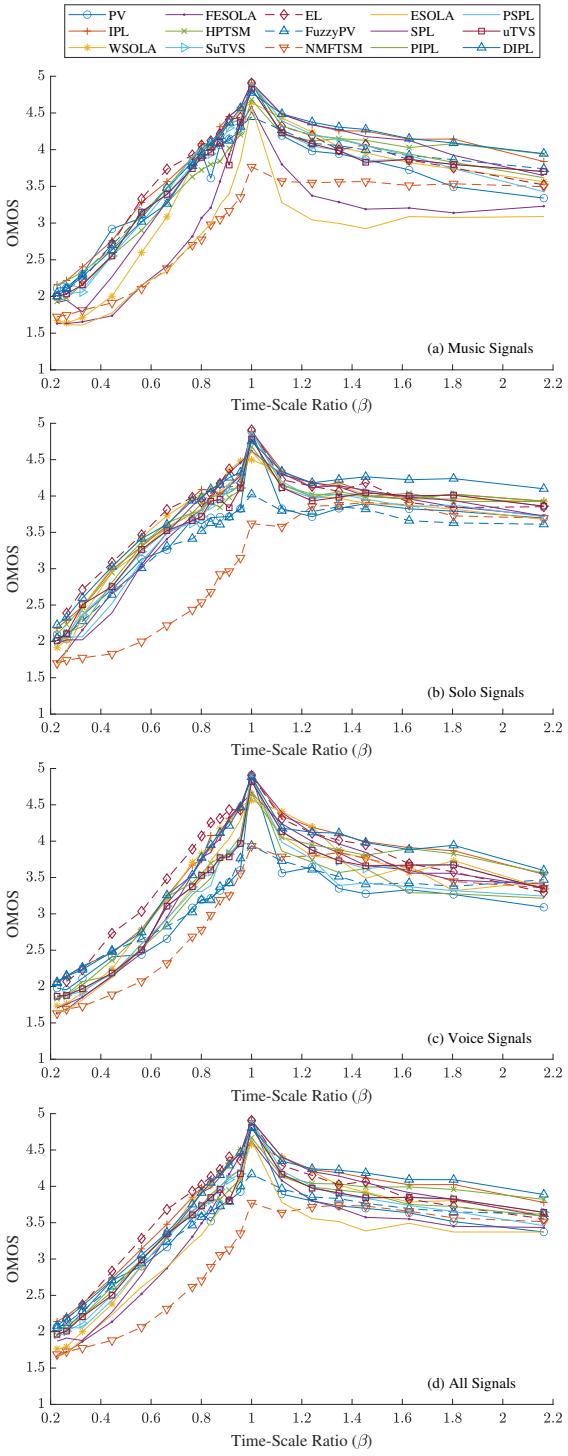

Figure 7.14: [Colour Online] Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined.

method. Secondly, expansion into alternative and deeper neural networks should also be considered. Initial testing resulted in a ρ_{te} of 0.71 for a random forest network using the hand-crafted features, while using blind data-driven features created by a CNN used as input to a fully connected network resulted in a ρ_{te} of 0.65.

7.6 Conclusion

An objective measure for time-scaled audio was proposed with performance superior to most subjective listeners. The measure used hand-crafted features and a fully connected network to predict subjective mean opinion scores. PEAQ Basic and Advanced features were used in addition to nine novel features specific to TSM artefacts. Six methods of alignment were explored, with interpolation of the magnitude spectrum to the duration of the test signal giving the best performance, achieving a mean RMSE of 0.490 and a mean PCC of 0.864. Using the proposed method to evaluate algorithms, it was found that Elastique gave the highest objective quality for voice signals, while the Identity Phase-Locking Phase Vocoder variants gave the highest objective quality for music and solo instrument signals as well as the best overall performance. Future work includes optimisation of feature generation, exploration of other network structures and evaluation of additional TSM algorithms.

Chapter 8

Deep Learning-Based Single-Ended Objective Quality Measures for Time-Scale Modified Audio

8.1 Introduction

The objective measure proposed in Chapter 7 (Roberts and Paliwal, 2020a) requires reference and test signals, and additional interpolation to align low-bandwidth representations of the signals. In this chapter, multiple single-ended objective measures of quality for audio processed with TSM are proposed. A convolutional or recurrent neural network front-end generates data-driven features, while a Fully-Connected Neural Network (FCNN) back-end predicts the overall quality. The measures are trained using the dataset of Chapter 6 (Roberts and Paliwal, 2020b), referred to as TSMDB from this point.

Subjective evaluation, such as BS.1284 ITU-T (2019), is the gold standard for evaluating quality of speech and audio processing. Participants are asked to rate the processing quality of audio files, often using ratings of Bad, Poor, Fair, Good and Excellent that map linearly to the interval [1,5]. Opinion scores are then averaged, giving a Mean Opinion Score (MOS) per file. This process however is lengthy and expensive. Consequently, many objective measures of quality have been proposed to predict MOS.

Objective measures can be classified into double-ended (DE) (invasive) and single-ended (SE) (non-invasive) methods. The former calculates differences between reference and processed signal pairs, while the latter operates solely on the processed signal. This allows non-invasive measures to be used in a variety of use cases such as testing of in-service real-time systems using multiple tests through a signal path, as in (Kim and Tarraf, 2007) and (Falk and Chan, 2006). SE measures have seen considerable use for speech quality (Falk and Chan (2006), Malfait et al. (2006), Sharma et al. (2016) and (Gamper et al., 2019)), while audio quality measures such as Thiede et al. (2000), Beerends et al. (2013), Huber and Kollmeier (2006) and Chinen et al. (2020) are DE. Additionally, some SE measures have been trained to DE measures, including Fu et al. (2018) and Catellier and Voran (2020).

SE methods are often compared to baseline DE measures such as Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001) and Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al., 2000). However, there is no standard for objective quality of TSM, with this dissertation forming the bulk of the work. The total published research is found in the following papers. Fierro and Välimäki (2020) published preliminary work towards an objective measure by using linear regression

of the mean squared error of transient, tonal and noise energy deviations to predict the Subjective Mean Opinion Score (SMOS). Initial measures were presented in Chapter 6 (Roberts and Paliwal, 2020b) and formalised in Chapter 7 (Roberts and Paliwal, 2020a), which extended PEAQ with additional features and explored synchronisation of reference and time-scale processed signals.

The DE method of Chapter 7, referred to as OMOQDE from this point, considered six methods of signal alignment before calculation of PEAQ features — in addition to hand-crafted features specific to the artefacts of time-scaled signals. Formulated as a regression problem, an FCNN was used to predict the MOS targets of the TSMDB. Alignment of reference and processed signals was achieved by interpolating the reference magnitude spectrum to the length of the processed spectrum before feature extraction. OMOQDE performance was improved by including reference files during training. Baseline performance was obtained by retraining the PEAQ Basic FCNN to Subjective Mean Opinion Score (SMOS) values. OMOQDE achieved an average Pearson Correlation Coefficient (PCC) ($\overline{\rho}$) of 0.864 and an average Root Mean Square Error (RMSE) loss ($\overline{\mathcal{L}}$) of 0.490 using the MOS range of 1-5, for the training, validation, and test sets. OMOQDE was able to resolve statistically significant differences in mean quality between TSM methods of 0.1 MOS. A distance measure that penalised overfitting was used to select the ideal network, see Section 7.2.4 above or the end of Section 8.2 below.

OMOQDE was trained using the TSMDB, which contains a training subset of 5280 files and a testing subset of 240 files. Traditional and state-of-the-art TSM methods were used in the training subset, with newer esoteric methods used in the testing subset. This resulted in no overlap between the TSM methods, time-scale ratios, or reference

signals. The TSM methods included:

- Phase Vocoder (PV) (Portnoff, 1976)
- Identity Phase-Locking Phase Vocoder (IPL) and Scaled Phase-Locking Phase Vocoder (SPL) (Laroche and Dolson, 1999)
- Waveform Similarity Overlap Add (WSOLA) (Verhelst and Roelands, 1993)
- Fuzzy Epoch Synchronous Overlap-Add (FESOLA) (Roberts and Paliwal, 2019)
- Harmonic Percussive Separation Time-Scale Modification (HPTSM) (Driedger et al., 2014)
- Mel-Scale Sub-band Modelling (uTVS) (Sharma et al., 2017) and the version used in subjective testing ($\overline{\text{uTVS}}$)
- Elastique (Zplane Development)
- Phase Vocoder using fuzzy classification of bins (FuzzyPV) (Damskägg and Välimäki, 2017)
- Non-Negative Matrix Factorisation Time-Scale Modification (NMFTSM) (Roma et al., 2019)
- PhaVoRIT ($\overline{\text{IPL}}$ and $\overline{\text{SPL}}$) (Karrer et al., 2006)
- Epoch Synchronous Overlap-Add (ESOLA) (Rudresh et al., 2018).
- IPL implementation of (Driedger et al., 2014) (DIPL).

Quality labels were provided as MOS and median opinion scores, calculated before and after session normalisation in the interval [1,5]. The

scores were collated from 42,529 ratings by 263 participants in 633 sessions, with a minimum of seven ratings per file. All files in the dataset have a single channel, a sampling rate of 44.1kHz, and bit depth of 16 bits. Some reference files are stereo, and were converted to a single channel by summation and normalisation to the interval [-1,1].

The usefulness of OMOQDE would be improved if a reference signal was no longer required, as it removes the need for signal alignment between reference and test signals. It could also reduce estimation time through bypassing the slow feature extraction process.

Deep learning is often used in objective measures of quality. Convolutional Neural Networks (CNNs) LeCun et al. (2015) are commonly used on spatial domain tasks, such as image classification. They have also found use in speech and audio due to the spatio-temporal representation of short-time frequency analysis, as in Gamper et al. (2019). CNNs learn weights of convolutional kernels which are applied successively creating higher order representations of the signal. Recurrent Neural Networks (RNNs) differ from standard fully-connected networks through the inclusion of a memory cell and are suited to time-series data. In this paper, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014) are the used cell types. LSTM cells are controlled by three gates, input, output and forget, which determine what information is added or removed from the cell. GRU is a variant of LSTM that removes the output gate and has fewer parameters. Bidirectional Recurrent Neural Networks (Schuster and Paliwal, 1997), such as the Bidirectional Long Short-Term Memory (BLSTM) and Bidirectional Gated Recurrent Units (BGRU), extend RNNs with forward and backward passes over the time-series.

Introduced by Davis and Mermelstein (1980), Mel Frequency Cep-

stral Coefficients (MFCCs) have found extensive use as a lower bandwidth transformed signal representation in speech processing, as in Nicolson et al. (2018). MFCCs are computed by first estimating the periodogram of the short-time power spectrum. A bank of triangular-shaped filters spaced uniformly on the mel-scale is then applied, resulting in the energy of each filter. The logarithm of the filterbank energies is then taken, followed by a Discrete Cosine Transform to decorrelate the filterbank energies. Differential and acceleration coefficients are often used to give an indication of the dynamics of the MFCCs and are generally known as Deltas (D) and Delta-Deltas (D′).

The chapter is organised as follows: Section 8.2 presents the proposed OMOQSE methods; Section 8.3 presents network results as well as a comparison of TSM algorithms. Availability, future research and conclusions are presented in Sections 8.4, 8.5 and 8.6 respectively.

8.2 Method

The proposed measures begin with audio processing. Signals were prepared by normalising to the interval [-1,1] and trimming silence at the beginning and end of the signal. Silence was determined, according to ITU-T (2001a), as the first and last time the sum of four consecutive samples is greater than 0.0061. The magnitude spectrum ($|X|$), magnitude and phase spectra ($[|X|;\angle X]$), power spectrum ($|X|^2$), MFCCs, MFCCs and D ([MFCCs;D]), and MFCCs, D and D′ ([MFCCs;D;D′]), where $[\cdot\ ;\ \cdot]$ is concatenation, were tested during development. The magnitude, phase, and power spectra used a frame length of $N = 2048$ samples, an overlap of N/2 and a Hann window. MFCCs were of length 128, with D and D′ width nine from $t - 4$ to $t + 4$ with respect to the

current time-step. Overall or per frequency-bin standardisation of the input features was explored.

Due to the variable length of the input signal, truncating and duplicating the signal were explored. For the CNN, sequences were truncated to the overall minimum length (L), starting from a different random location in each epoch. During testing, the OMOS was averaged over 16 segments to capture more information of the processed signal, for a wider selection of input signals. Repeating the input signal to the duration of the longest signal was also considered for GRU-FT, however as LSTM and GRU operate sequentially on each frame, input signals were used in their entirety.

Prior to network training, target scores were scaled to the interval [0,1] using Equation 7.13.

The proposed CNN data structure is shown in Figure 8.1. It contains four convolution layers, of filter sizes 16, 32, 64 and 32, with batch normalisation and a 5x5 kernel for the first layer, and 3x3 for the remaining layers. The first two convolutional layers are followed by max pooling layers, with 2x2 kernels and 2x2 stride. After concatenation and 10% dropout, three fully-connected layers of output size 128 are used. The final layer has a single output. This results in 821,857 trainable parameters. Rectified linear unit (ReLU) activation is used throughout, except for the output layer, where the Sigmoid activation is used. Residual connections around the second and third fully-connected layers are used. Root Mean Square Error (RMSE) is used as the loss function. Features were concatenated in time-aligned input panes.

The proposed final-frame (FF) model for LSTM, BLSTM, GRU and BGRU networks can be seen in Figure 8.2. FF RNN models use back

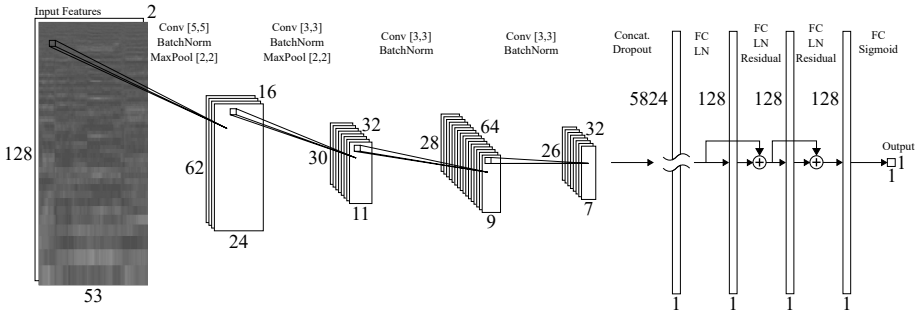

Figure 8.1: Proposed CNN dataflow. Kernel sizes in brackets, numbers denote layer size and number of channels, FC is a fully-connected layer, LN is layer normalisation, ReLU activation used unless specified.

propagation through time to learn from the error between the final output and the SMOS. The total feature dimension (D_F) is set by the concatenation of input features. For the proposed network using [MFCCs;D] features, D_F is 256. Two RNN layers were used with the memory layer size (D_H) set to the number of directions (n) multiplied by D_F. L is the sequence length and ranged from 53 to 2179 frames. An RNN architecture of many-to-one was used, with the final frame used as input to an FCNN after 10% dropout. The FCNN contained three layers of output size 256, 128 and 1, respectively. Layer normalisation and ReLU activation were used for layers 1 and 2, while Sigmoid activation was used for the output layer. This results in 16,370,945 trainable parameters. Again, RMSE is used as the loss function. Magnitude, Phase and Power spectra ($D_F = 1025, D_H = 512$) were also explored as input to this network.

The proposed frame-target (FT) model for GRU and BGRU networks (GRU-FT and BGRU-FT) can be seen in Figure 8.3. Two GRU layers of $D_H = 256$ with 10% dropout were used in a similar structure to the previous RNN. However, a single fully-connected layer with sigmoid activation reduces feature dimensionality to $L \times 1$. The network has

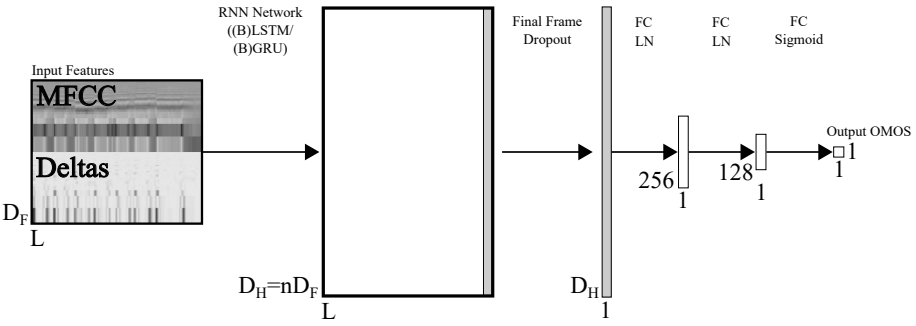

Figure 8.2: Proposed RNN FF dataflow. D_F is feature depth, D_H is Hidden Dimensions, n is the number of directions, numbers denote layer sizes, FC is a fully-connected layer, LN is layer normalisation, ReLU activation used unless specified.

1,972,737 trainable parameters. The Mean Square Error (MSE) between the target SMOS and each frame estimate is used as loss. Frame targets are averaged for the length of the sequence to calculate the OMOS. As this calculation is independent of training, median, minimum and maximum values of frame targets were also considered. Minimum frame targets were considered as quality evaluation of time-scaled signals is a degradation style analysis, where subjective quality is heavily influenced by the quality of the worst part of the signal. Due to the inverse correlation between time-scale ratio and SMOS for signals that have been slowed down, an inverse exponential relationship between the number of frames at the time-scale and the time-scale itself, possibly leading to difficulty in estimating quality for signals that have been sped up. The impact of this was explored by training on signals truncated to the minimum signal length and on signals repeated to the maximum signal length.

10% of the training dataset was reserved for validation. The CNN was trained for 100 epochs using a mini-batch size of 132, while RNNs were trained for 30 to 60 epochs with a mini-batch size of 48. A learning

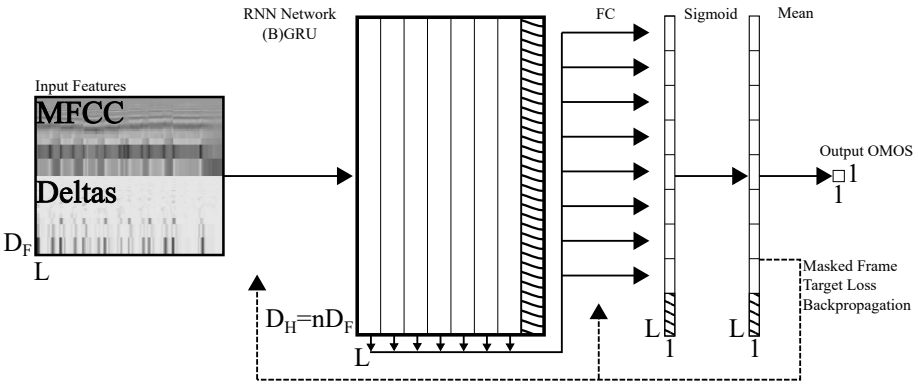

Figure 8.3: Proposed GRU FT network dataflow. D_F is feature depth, D_H is hidden dimensions, n is the number of directions, L is sequence length, numbers denote layer sizes, FC is a fully-connected layer and hashed sections are zero-padding to longest file in mini-batch.

rate of $1e^{-4}$ was used in most cases, with $1e^{-5}$ if network performance stopped improving within the first 10 epochs. AdamW (Loshchilov and Hutter, 2017) was used as the optimiser for all networks. Loss for back-propagation was calculated using estimates in the interval of [0,1]. Reported loss values ($\mathcal{L}$) were calculated using RMSE and estimates scaled back to the original interval of [1,5], for comparison with OMOQDE. As the prediction of opinion scores for novel TSM methods is the use case, early stopping based on validation loss was not used. The same scheme as Section 6.5 was used to select the optimal epoch.

An evaluation set of 6000 files, detailed in Chapter 7, was generated from the reference files in the test set. 20 new time-scales in the range of $0.22 < \beta < 2.2$, with all TSM methods listed in Section 8.1 used to process the reference files. During evaluation, averages do not include $\beta = 0.2257$ as the minimum for Elastique is $\beta = 0.25$, or $\beta = 1$ as it should result in a unity system. This is not always the case, and can be useful for determining method performance, but it has been excluded from the analysis.

8.3 Results

8.3.1 Network Performance

A wide range of testing and network configurations were considered during the development of the proposed measures. Network hyper-parameters were optimised through a systematic non-exhaustive search. Deterministic training of all networks was conducted using seeds from 0 to 29. The use of RMSE, Mean Square Error and Mean Absolute error were explored, with the best performing loss function used in each proposed method. Figure 8.4 shows the box plot distribution of the best $\mathcal{D}$ for each seed, where lower is better.

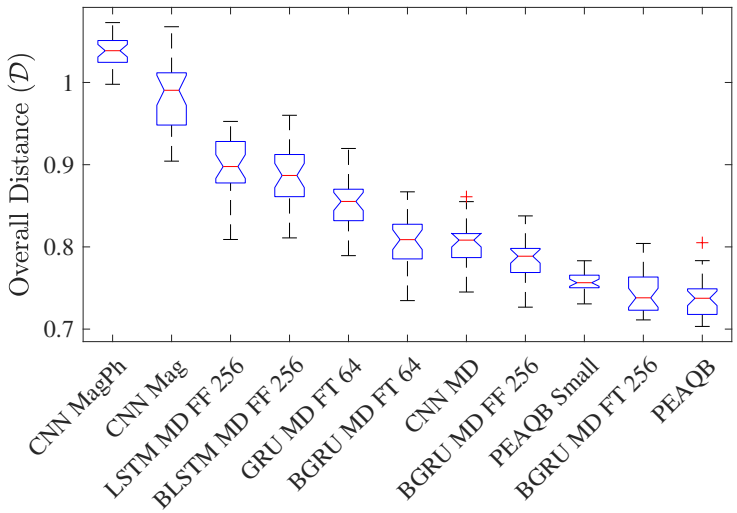

Figure 8.4: Box plot of best distance measure for 30 seeds of each network configuration, ordered by median $\mathcal{D}$. $|X|$ is denoted by Mag, $\angle X$ by Ph, [MFCCs;D] by MD and hidden size denoted by 64 or 256.

Median overall distance ($\widetilde{\mathcal{D}}$) and the best case $\boldsymbol{\mathcal{D}}$ with associated $\overline{\boldsymbol{\mathcal{L}}}$, $\Delta\boldsymbol{\mathcal{L}}$, $\overline{\boldsymbol{\rho}}$ and $\Delta\boldsymbol{\rho}$ values can be found in Table 8.1, along with ($\mathcal{L}_{te}$) and (ρ_{te}). While the improvement in performance appears linear in Figure 8.4, many network configurations have not been included. Most

networks trained with [MFCCs;D] achieved $0.55 < \mathcal{L}_{te} < 0.67$, with only BGRU-FT achieving $\mathcal{L}_{te} > 0.68$ or $\mathcal{D} < 0.72$. This appears to be the $\mathcal{L}_{te}$ and $\mathcal{D}$ limit for these network configurations and input features, even with ρ_{tr} approaching 1 when allowed to over-train.

The results in Table 8.1 can be summarised as follows. The proposed CNN achieved a best $\overline{\boldsymbol{\mathcal{L}}}$ of 0.608, $\overline{\boldsymbol{\rho}}$ of 0.771, $\mathcal{L}_{te}$ of 0.801 and ρ_{te} of 0.637 placing it at the 74th and 32nd percentiles of subjective sessions for $\overline{\boldsymbol{\mathcal{L}}}$ and $\overline{\boldsymbol{\rho}}$ respectively. The proposed BGRU-FT network achieved a best $\overline{\boldsymbol{\mathcal{L}}}$ of 0.576, $\overline{\boldsymbol{\rho}}$ of 0.794, $\mathcal{L}_{te}$ of 0.762 and ρ_{te} of 0.682 placing it at the 84th and 39th percentiles of subjective sessions for $\overline{\boldsymbol{\mathcal{L}}}$ and $\overline{\boldsymbol{\rho}}$ respectively.

To given an indication of what the networks may be learning, correlation between OMOQSE or OMOS and OMOQDE features was calculated for CNN and BGRU-FT networks. No significant correlation was found with maximum correlations of 0.210 and 0.206 for CNN and BGRU-FT respectively.

Several trends were seen across testing. Networks trained using [MFCCs;D] features out-performed those trained using [MFCCs;D;D′], as well as solely MFCCs, magnitude spectra, magnitude and phase spectra, and the power spectrum. In all cases, magnitude only features out-performed combined magnitude and phase features. Decreased performance due to the inclusion of phase is likely due to its noise-like quality. Results for networks trained on the power spectrum are not shown in plots to increase comprehension. Improved performance was found using MFCCs generated with Librosa over TorchAudio, with identical settings. For RNNs, FT measures outperformed FF measures, GRU outperformed LSTM, and bidirectional networks generally outperformed single direction networks for the same input features and network size. As such, RNN analysis will focus on BGRU-FT, alongside analysis of

Table 8.1: Test Loss ($\boldsymbol{\mathcal{L}_{te}}$) and PCC ($\boldsymbol{\rho_{te}}$), mean RSME loss ($\overline{\boldsymbol{\mathcal{L}}}$) and range ($\boldsymbol{\Delta\mathcal{L}}$), mean PCC ($\overline{\boldsymbol{\rho}}$) and range ($\boldsymbol{\Delta\rho}$), median overall distance ($\widetilde{\boldsymbol{\mathcal{D}}}$) and minimum overall distance ($\min(\boldsymbol{\mathcal{D}})$). Best single-ended results in bold.

Ended	**Network**	**Features**	**Hidden**	$\boldsymbol{\mathcal{L}_{te}}$	$\boldsymbol{\rho_{te}}$	$\overline{\boldsymbol{\mathcal{L}}}$	$\boldsymbol{\Delta\mathcal{L}}$	$\overline{\boldsymbol{\rho}}$	$\boldsymbol{\Delta\rho}$	$\widetilde{\boldsymbol{\mathcal{D}}}$	$\min(\boldsymbol{\mathcal{D}})$
SE	BLSTM-FF	$\lvert X\rvert^2$	512	1.123	0.244	1.009	0.194	0.163	0.220	1.364	1.344
SE	LLSTM-FF	$\lvert X\rvert^2$	512	1.064	0.262	0.989	**0.135**	0.171	0.251	1.350	1.322
SE	CNN	$[\lvert X\rvert;\angle X]$	-	0.942	0.484	0.850	0.188	0.523	0.099	1.039	0.998
SE	CNN	$\lvert X\rvert$	-	0.944	0.553	0.745	0.339	0.674	0.205	0.991	0.904
SE	LSTM-FF	[MFCCs;D]	256	0.854	0.581	0.663	0.295	0.720	0.221	0.898	0.809
SE	BLSTM-FF	[MFCCs;D]	256	0.849	0.581	0.670	0.282	0.711	0.214	0.887	0.811
SE	GRU-FT	[MFCCs;D]	64	0.820	0.649	0.699	0.188	0.701	**0.097**	0.855	0.789
SE	BGRU-FT	[MFCCs;D]	64	0.778	0.675	0.611	0.287	0.770	0.179	0.809	0.735
SE	CNN	[MFCCs;D]	-	0.801	0.637	0.608	0.301	0.771	0.206	0.808	0.745
SE	BGRU-FF	[MFCCs;D]	256	0.784	0.667	0.622	0.248	0.762	0.153	0.789	0.727
DE	FCNN	PEAQB	3	0.691	0.749	0.668	0.054	0.719	0.075	0.756	0.731
SE	BGRU-FT	[MFCCs;D]	256	**0.762**	**0.682**	**0.576**	0.307	**0.794**	0.192	**0.738**	**0.711**
DE	FCNN	PEAQB	128	0.704	0.742	0.650	0.091	0.748	0.009	0.738	0.703
DE	FCNN	To Test Incl Ref	128	0.550	0.852	0.490	0.101	0.864	0.030	0.600	0.519

CNN performance.

The CNN improved significantly through the use of [MFCCs;D] over $|X|$, $\angle X$ and $|X|^2$, with similar performance to FF RNNs. Normalisation of input spectra reduced network performance. Small gains were found through optimising the kernel size, however the maximum kernel size was limited by the length of the shortest file. Repeating signals to the length of the longest files decreased network performance, as did using a combination of repeating or truncating to 500 or 1000 frames. The CNN predicts across most of the OMOS range, shown in Figure 8.5, and achieves a correlation of 0.564872 with the OMOQDE OMOS. Loss and correlation can be found in Table 8.1.

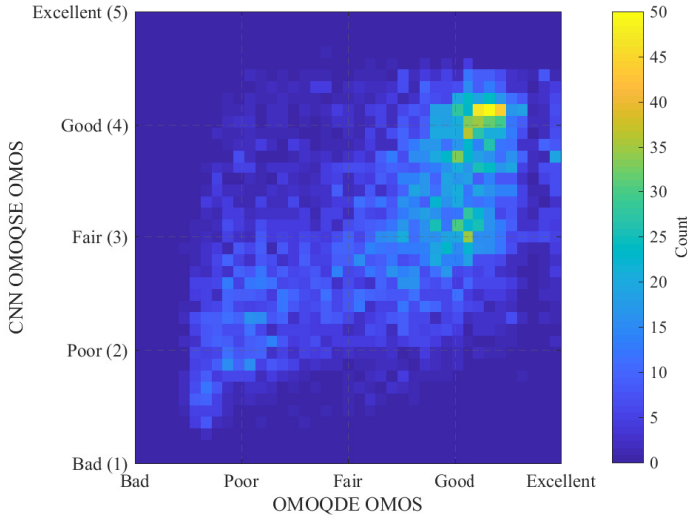

Figure 8.5: [Colour Online] OMOS confusion matrix for CNN OMOQSE and OMOQDE.

Figure 8.6 shows the prediction of the CNN for the test subset overlaid on the training subset. A lack of lower extension in test set prediction is clearly visible.

The BGRU-FT was found to give the best performance of the tested

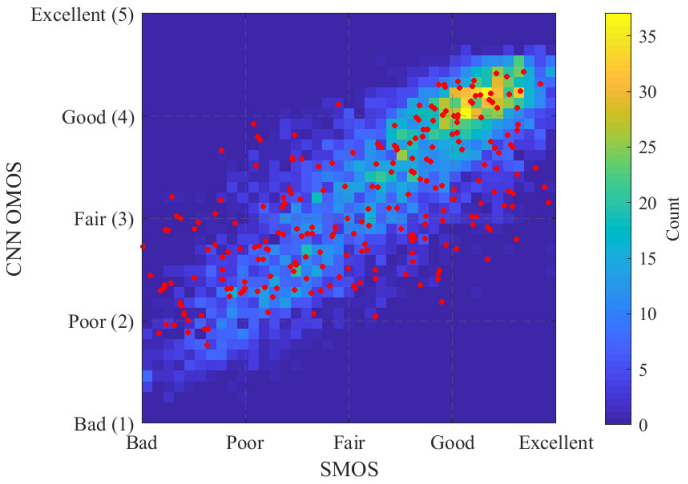

Figure 8.6: [Colour Online] Training subset confusion matrix for CNN OMOQSE and SMOS. Test set shown as red dots.

SE networks according the distance measure, and gives similar performance to OMOQDE trained using PEAQ Basic features. $\overline{\boldsymbol{\mathcal{L}}}$ and $\overline{\boldsymbol{\rho}}$ are improved over PEAQB networks, despite worse $\mathcal{L}_{te}$ and ρ_{te}, resulting in larger $\Delta\boldsymbol{\mathcal{L}}$ and $\Delta\boldsymbol{\rho}$ values, shown in Table 8.1. When collapsing estimated frame targets, no significant difference was found between mean or median of predictions, while selecting the minimum or maximum prediction reduced performance. A short-coming of most BGRU-FT networks trained is the lack of predictions for $OMOS > 4$, which can be seen in Figure 8.7, where the correlation is 0.549. A hidden size of 256 outperformed 64, 128 and 512 sizes, with 10% dropout outperforming 0%, 25% and 50%. Including D′ was found to reduce performance, as did increasing the number of MFCCs to 256. Multiple fully-connected layers were also explored, but did not improve performance. FT RNNs slightly improved performance over FF RNNs, with FF improvements following BGRU-FT results, with the best FF network shown in Figure 8.4 and Table 8.1.

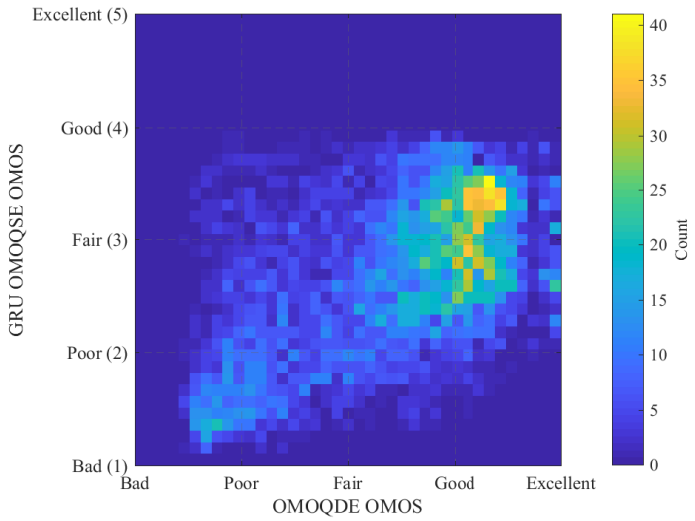

Figure 8.7: [Colour Online] OMOS confusion matrix for BGRU-FT OMOQSE and OMOQDE.

Figure 8.8 shows the prediction of the BGRU-FT network for the test subset overlaid on the training subset. The test set shows greater lower extension than the CNN, but still shows a significant spread in prediction away from the true SMOS.

Experiments showed that using truncated random segments with BGRU-FT reduced performance, as did extending signals through repetition. Repeating input for the CNN also reduced performance. While the number of frames for $\beta << 1$ in the training subset is significantly greater than for $\beta >> 1$, the number frames is relatively uniform for $2 \lesssim SMOS \lesssim 4.5$, see Figure 8.9. The reduced number of frames for $1 \leq SMOS \lesssim 1.5$ and $4.5 \lesssim SMOS \leq 5$ may also impact the estimation at outermost OMOS, as seen in Figure 8.7. Surprisingly, although truncated segments are used as input to the CNN, the estimates show a wide spread in Figure 8.5.

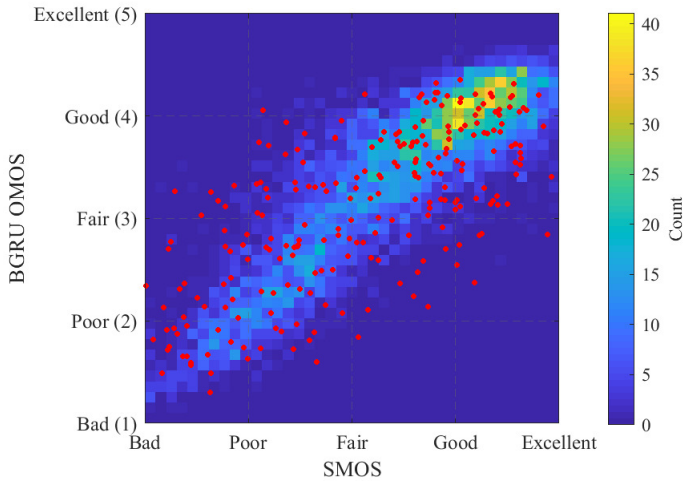

Figure 8.8: [Colour Online] Training subset confusion matrix for BGRU-FT OMOQSE and SMOS. Test set shown as red dots.

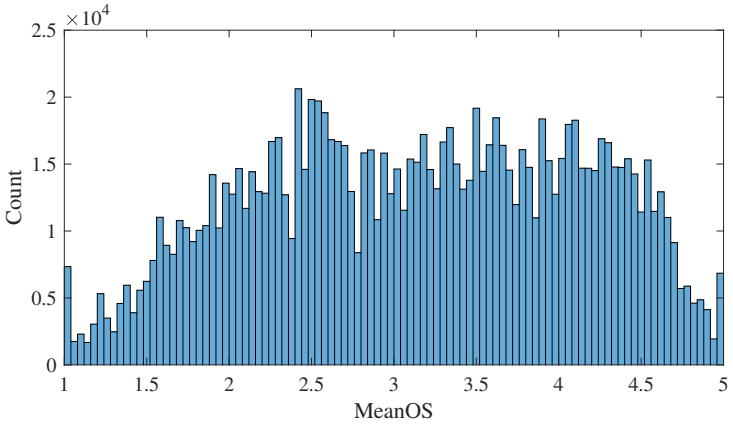

Figure 8.9: Distribution of frames per MOS in training set.

8.3.2 TSM Algorithm Evaluation

In this section, TSM methods are evaluated using the aforementioned evaluation set. Tables 8.2 and 8.3 show average OMOS for each signal class per TSM method ordered by overall mean OMOS, Figure 8.10 and 8.12 show average OMOS per TSM method and β and Figure 8.11 and 8.13 show TSM methods for which differences in mean are statis-

tically significant. As in Chapter 7, all results for $\beta = 1$ and $\beta < 0.25$ were excluded from averaging calculations forming Tables 8.2 and 8.3, as time-scaling is applied at $\beta \neq 1$, and the minimum β available for Elastique is 0.25. Common trends are presented, followed by CNN and then BGRU-FT analysis.

Estimation of signals time-scaled using NMFTSM was particularly challenging for all networks. This is likely due to novel artefacts described by Roma et al. (2019) and the SMOS distribution skewed towards low scores. This provides a challenge for network design as novel TSM methods may not have similar artefacts or SMOS distributions to those in the training set. However, the relative rating of Elastique and FuzzyPV to other TSM methods follows that of subjective testing. As suggested by network $\mathcal{L}_{te}$ and ρ_{te}, only a general sense of TSM quality is obtained. Small details, such as the reduced quality of uTVS used in subjective testing at $\beta \approx 1$, are not visible. The networks have also not learnt the non-linearity of SMOS as a function of β, continuing to increase for $\beta > 1$, seen in Figure 8.10 and 8.12. The uniform quality of methods at $\beta = 1$ is however visible, as is the reduction in TSM quality for $\beta < 1$.

For musical files, Figure 8.10(a), the CNN differentiates between frequency and time-domain methods, where quality rapidly falls for time-domain methods when $\beta < 1$. WSOLA fairs the best of the time-domain methods, diverging from frequency domain methods for $\beta < 0.8$. The relative improvement in PV quality is also visible for $\beta < 0.5$, and the slower falloff of Elastique. When averaged, the CNN rates uTVS and subjective uTVS highest followed by Elastique. For solo files, (Figure 8.10(b), HPTSM exceeds other methods for $\beta < 1$. This class is the only evaluation where the highest OMOS at $\beta = 1$. HPTSM has

the highest mean OMOS, followed by PhaVoRIT IPL, both uTVS methods, Elastique, and WSOLA, as shown in Table 8.2. Voice file OMOS, Figure 8.10(c), is comparatively low for time-domain methods, which is unexpected, as speech is often scaled well by time-domain methods. The high quality of NMFTSM is also unexpected based on subjective results in Chapter 6. Elastique has the highest mean OMOS, followed by NMFTSM, SPL and IPL. All other methods gave similar averaged quality.

Table 8.2: Mean OMOS for each class of file and overall result for the proposed CNN OMOQ. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$.

TSM Method	Music	Solo	Voice	Overall
ESOLA	2.291	3.248	2.966	2.781
FESOLA	2.424	3.209	2.938	2.814
PV	3.318	3.202	2.793	3.126
WSOLA	3.045	3.416	3.020	3.149
NMFTSM	3.053	3.370	3.082	3.157
FuzzyPV	3.290	3.396	2.886	3.200
SPL	3.259	3.297	3.064	3.212
Driedger's IPL	3.299	3.343	2.982	3.217
Phavorit SPL	3.327	3.375	2.948	3.227
Phavorit IPL	3.335	3.441	2.879	3.230
IPL	3.294	3.372	3.053	3.245
HPTSM	3.325	**3.553**	2.925	3.274
Subjective uTVS	3.460	3.424	2.988	3.307
uTVS	**3.469**	3.435	2.999	3.318
Elastique	3.445	3.419	**3.104**	**3.335**

For all CNN OMOS, Elastique has the highest average rating followed by both uTVS methods and HPTSM. The overall means can be seen in Figure 8.10(d). The best five methods are separated by < 0.1 OMOS with a maximum difference of 0.554 for all methods. The overall low quality of FuzzyPV is unexpected given that it builds on IPL, however further analysis is required to determine if the difference statistically significant. A two-sample t-test ($\alpha = 0.05$) of all OMOS shows the null

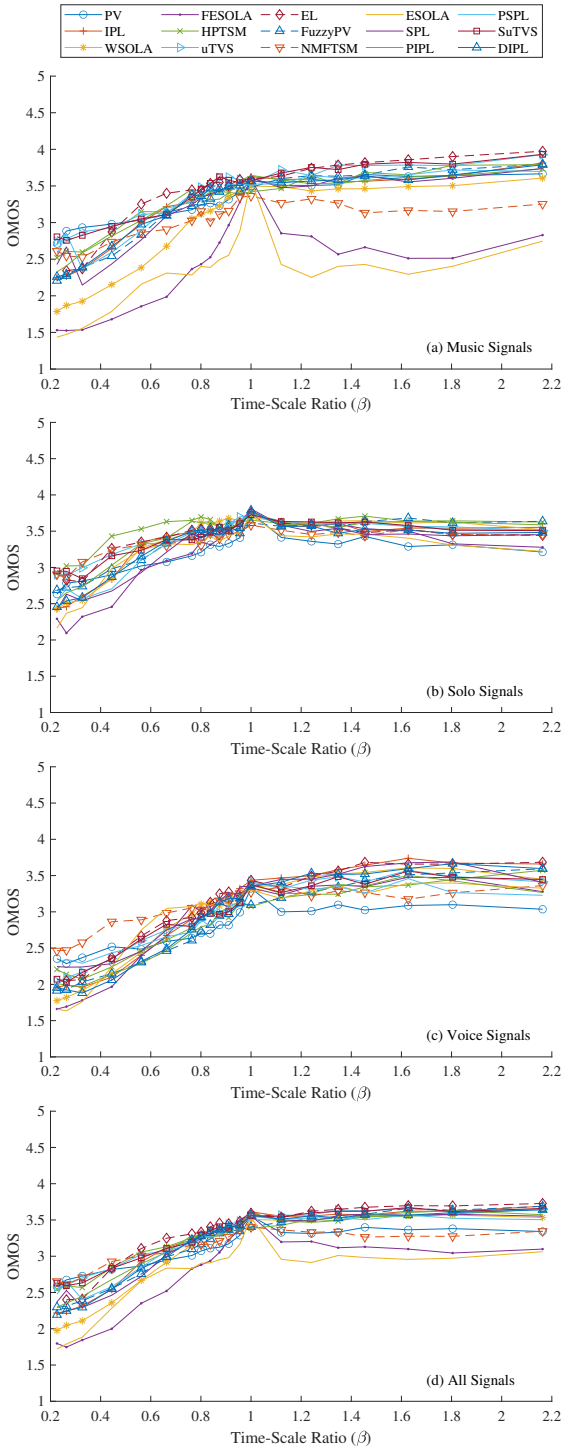

Figure 8.10: [Colour Online] CNN estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined.

hypothesis, TSM methods having equal means, to be rejected in almost all cases when the absolute difference of mean OMOS is greater than 0.098. Figure 8.11 shows p-values for the t-tests that were unable to reject equal means.

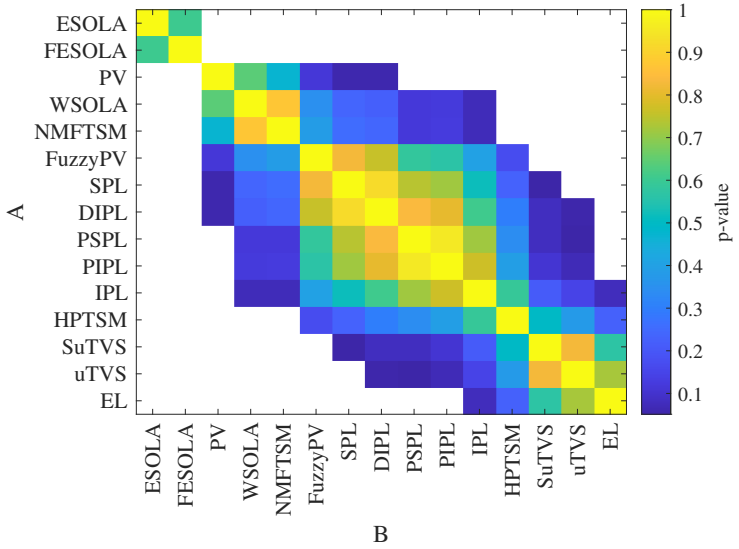

Figure 8.11: [Colour Online] Masked two-sample t-test for all CNN OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white.

BGRU-FT OMOS shows the most variance for music files, Figure 8.12(a). Again, time-domain methods rate lower. FuzzyPV is rated highest, followed by uTVS and Elastique. For multiple TSM methods, OMOS continues to increase for $\beta > 1$. By combining this information with the improvement when D is included as an input feature, we theorise that BGRU-FT is learning a relationship between SMOS and the velocity and duration between D events. As β increases, the time between sound events decreases and the attack portion of the energy envelope becomes sharper. For solo files, Figure 8.12(b), there is very little variance between methods, with a maximum difference $\approx$0.5 OMOS for $\beta = 0.2257$. Solo files have the highest overall OMOS of

the three classes, which is consistent with subjective findings. HPTSM has the highest mean OMOS, followed by uTVS and FuzzyPV. Voice file OMOS, Figure 8.12(c), shows the lowest TSM quality of the three classes with a continued increase in OMOS for $\beta > 1$ across all TSM methods. NMFTSM has the highest mean OMOS, which is unexpected based on Chapter 6. Elastique is next highest followed by uTVS methods and ESOLA. The high quality of ESOLA is expected as the method was designed for TSM of speech.

Table 8.3: Mean OMOS for each class of file and overall result for the proposed BGRU-FT network. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$.

TSM Method	Music	Solo	Voice	Overall
ESOLA	2.378	2.925	2.503	2.580
FESOLA	2.511	2.947	2.468	2.629
PV	2.723	2.891	2.323	2.653
SPL	2.711	2.917	2.350	2.664
IPL	2.764	2.917	2.329	2.680
Phaovrit SPL	2.733	2.943	2.345	2.680
Phavorit IPL	2.764	2.976	2.334	2.699
WSOLA	2.741	2.978	2.472	2.732
HPTSM	2.699	**3.035**	2.492	2.738
Drideger's IPL	2.787	2.978	2.443	2.741
NMFTSM	2.720	2.988	**2.591**	2.762
Subjective uTVS	2.890	2.983	2.526	2.809
Elastique	2.899	2.988	2.544	2.819
uTVS	2.901	3.005	2.532	2.822
FuzzyPV	**3.016**	3.011	2.444	**2.843**

For overall OMOS, Figure 8.10(d), FuzzyPV has the highest average rating followed by uTVS methods and Elastique. The ordered ranking of methods is close to expected, with only NMFTSM ranking unexpectedly. This is possibly due to the method retaining the shape of percussive elements during time-scaling. The six best methods are separated by < 0.102 OMOS, with a maximum difference of 0.263 for all methods.

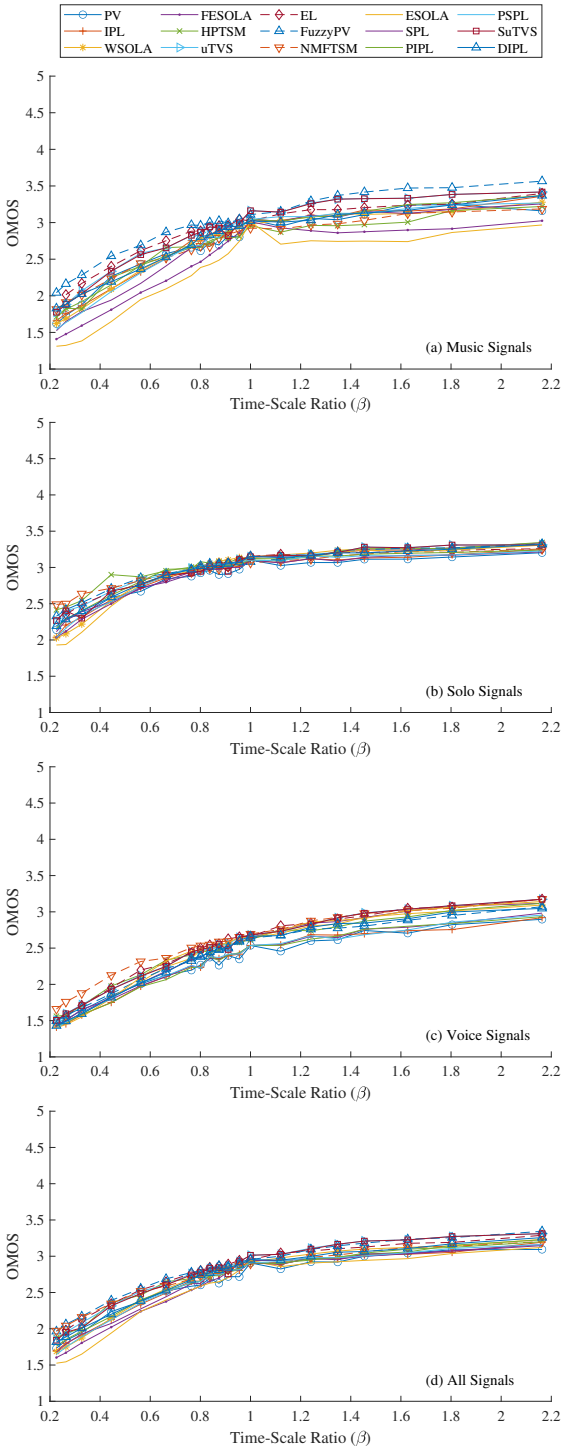

Figure 8.12: [Colour Online] BGRU-FT estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals and (d) All signals combined.

A two-sample t-test analysis ($\alpha = 0.05$) of all OMOS shows the null hypothesis of equal means to be rejected in almost all cases when the absolute difference of mean OMOS is greater than 0.098. Figure 8.13 shows p-values for the t-tests that were unable to reject equal means.

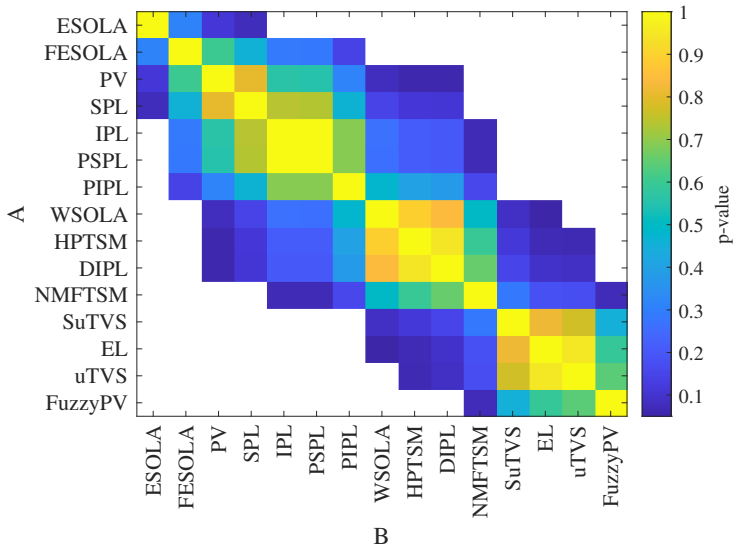

Figure 8.13: [Colour Online] Masked two-sample t-test for all BGRU-FT OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white.

The OMOQDE took approximately 15 hours to evaluate the 6000 files of the evaluation set (approximately 7 hours of audio), while the proposed networks took approximately 15 minutes on a system with a Xeon E5-2630 and an NVIDIA GTX1080. The majority of this improvement is due to the removal of time-frequency spreading when calculating PEAQ features. The elimination of alignment between reference and test signals is also beneficial as it removes an additional temporal manipulation before feature calculation. While OMOQDE is a more accurate estimate of time-scaling quality, the proposed OMOQSE measures give a fast relative quality assessment, and provide a platform for future SE objective measures.

8.4 Availability

The proposed CNN and BGRU-FT tools are available for free from github.com/zygurt/TSM. This includes python scripts for feature generation, the proposed methods implemented in PyTorch and evaluation methods. A bash script is also included to simplify use.

8.5 Future Research

This study shows promise in non-invasive evaluation of the quality of TSM methods. However, improvements can be made through input feature selection and exploring the use of phase derivatives or instantaneous frequency. Generalisation to unseen TSM methods and sound sources is also an area for future research. More research needs to be conducted regarding duration invariant transformations that limit the networks ability to learn simple relationships such as the duration of musical events within the signal to SMOS. Additional attention could also be given to network architectures, such as Transformer networks (Vaswani et al., 2017). Pre-training using a large task related dataset could also be explored.

8.6 Conclusion

Two single-ended objective measures for time-scaled audio were proposed with performance matching that of simple OMOQDE measures with reduced processing time. CNN and BGRU-FT network architectures were used to generate data-driven features from [MFCCs;D] inputs, which were then fed to an FCNN. The networks were trained to

the SMOS of the TSMDB. The CNN achieved an $\overline{\mathcal{L}}$ of 0.608 and $\overline{\rho}$ of 0.771, while BGRU-FT achieved an $\overline{\mathcal{L}}$ of 0.576 and $\overline{\rho}$ of 0.794. The proposed measures were used to evaluate TSM methods, with estimates consistent with relative quality found in subjective testing.

Part IV

Conclusion

Chapter 9

Dissertation summary: Conclusions and Future Research

9.1 Chapter 5.1: Stereo Time-Scale Modification

In this chapter we explored the shortcomings of most time-scale modification methods when processing stereo signals. Following this, two methods of maintaining the stereo image were detailed. These methods used either pre-and post-processing of the entire signal or frame, in order to maintain the stereo phase coherence. Sum and difference transformations of the stereo signal were calculated before processing and then transformed back after TSM processing. As a result, the stereo field was maintained through a reduction in the loss of stereo phase coherence. The proposed methods matched previous methods, producing

a high quality stereo output and greatly improved quality over the independent channel processing method. It was also shown that the method allowed for simple implementation around previous TSM methods, and is suitable for use with frequency and time-domain TSM methods.

9.2 Chapter 5.2: Fuzzy Epoch-Synchronous Overlap-Add

In this chapter, a modification to ESOLA was proposed. Instead of simply aligning the next epoch, cross-correlation between time-smeared epochs was used to calculate the optimal overlap location. This reduced distortion and artefacts, particularly during changes in pitch. The proposed method was tested against well known TSM methods, and was found to be preferred over ESOLA, and give similar quality to other TSM methods when processing speech signals. It was also shown that the proposed method can effectively scale monophonic instrument signals with strong fundamental frequencies.

9.3 Chapter 6: A Time-Scale Modification Dataset with Subjective Quality Labels

In chapter 6 the creation, subjective evaluation and analysis of a dataset is detailed. Six TSM methods were used to process 88 reference files at 10 time-scales, resulting in 5,280 processed signals for a training subset. Three additional methods at four random time-scales processed an additional 20 reference files, resulting in 240 signals for a testing subset. Large scale subjective testing of the dataset was detailed, with 42,529 ratings collected from 633 sessions. Statistical analysis was pre-

sented, validating the resulting mean and median opinion scores. The dataset was then used in the development of a proposed objective measure of quality for time-scaled audio. Preliminary results for the objective measure of quality were presented, achieving a Root Mean Square Error (RMSE) loss of 0.668 and Pearson Correlation Coefficient (PCC) of 0.719 for the test set. This dataset was then used in the following two chapters.

9.4 Chapter 7: An Objective Measure of Quality for Time-Scale Modification of Audio

In this chapter, the first effective objective measure of quality for time-scaled audio was proposed. Hand-crafted features and a fully connected network were used to predict subjective mean opinion scores from Chapter 6. Perceptual Evaluation of Audio Quality Basic and Advanced features were used in addition to nine TSM specific features. Six methods of signal alignment were proposed, with network training results showing that interpolation of the magnitude spectrum to the duration of the test signal gave the best performance. The proposed network achieved a mean RMSE of 0.490 and a mean PCC of 0.864. These results placed the objective measure at the 97th and 82nd percentiles of subjective sessions for RMSE and PCC, respectively.

The proposed measure was then used to evaluate time-scale modification methods, with Elastique giving the highest objective quality for voice signals and the Identity Phase-Locking Phase Vocoder variants giving the highest objective quality for music and solo instrument signals as well as the best overall performance.

9.5 Chapter 8: Deep Learning-Based Single-Ended Objective Quality Measures for Time-Scale Modified Audio

In Chapter 8, two single-ended objective measures for time-scaled audio were proposed. The performance of these measures was equivalent to simple OMOQDE measures with reduced processing time. Data driven features extracted from MFCC and Delta transformations of processed signals, using either a Convolutional Neural Network or Bidirectional Gated Recurrent Unit network architectures, were used as input to a fully-connected neural network. The networks were trained to the subjective mean opinion scores of Chapter 6. The CNN achieved a mean RMSE loss of 0.608 and mean PCC of 0.771, while BGRU-FT achieved a mean RMSE loss of 0.576 and mean PCC of 0.794. Finally, the proposed measures were used to evaluate TSM methods, with estimates consistent with relative quality comparisons found in subjective testing.

9.6 Future Research

There are many direction for future work that build on this dissertation. Some of this work is in the application of the proposed measures, while some is in the extension and improvement of the work.

Based on this research a number of recommendations can be presented to assist future subjective testing. The number of files presented in each session should be kept as small as practically possible, particularly for inexperienced listeners. We found 60 to 80 files to be optimal in terms of participation and completion of sessions. The ease of use of the testing software should also be considered, with recruiting partici-

pants became easier once using the online solution. Additionally, using the level of impairment qualifier could be a better choice than Bad to Excellent due to the large artefacts present in TSM.

While a large number of time-scale modification algorithms were evaluated using the objective measure, there are still many methods that have not been tested, such as the shape invariant time-scale modification of Quatieri and McAulay (1992); the approach to transients of Röbel (2003); the non-linear time-scaling of Ravelli et al. (2005); PVSOLA of Moinet and Dutoit (2011); GaborTSM of Ottosen and Dörfler (2017); and Phase Vocoder Done Right of Průša and Holighaus (2017). This is particularly true of commercial methods, which have been rarely included in subjective evaluations. These methods include IRCAM Lab TS (ircamLab), Melodyne 5 (Melodyne), izotope Time & Pitch (izotope), Paul's Extreme Sound Stretch (PaulStretch). The measures could also be used to optimise parameters and choices during implementation of a specific method.

The work of Chapter 7 could be furthered through optimisation of feature calculation. Particularly the time and frequency smearing within the PEAQ algorithm. An efficient method has been proposed by Kabal et al. (2002), however it was not implemented to maintain consistency with ITU-T (2001a). This change would drastically reduce the computation time, and increase the usability of the measure. Additionally, the features could be applied to an alternative network architecture. Initial testing of an un-tuned Random Forrest Network achieved a test PCC of 0.71. The features of Chapter 7 could also be explored further. The final calculation for most features is time or frequency domain averaging of a vector. Inclusion of the standard deviation of each vector may improve performance. Alternatively, a measure may be developed using

the feature vectors directly. This measure would need to account for the variable feature length between files, and different length features. Finally, additional features could be developed using transformations such as the scale invariant Fast Mellin Transform of Bertrand et al. (1990).

The work of Chapter 8 shows promise for non-invasive evaluation of time-scale modification quality. However, improvements could be found through input feature selection and exploring the use of phase derivatives or instantaneous frequency. Improving the network generalisation for unseen TSM methods is also a potential area for future research. Additionally, the Fast Mellin Transform (Bertrand et al., 1990) may be beneficial as a transformation for initial input features to limit the networks ability to learn simple relationships such as the duration of musical events within the signal to SMOS. Finally, attention could also be given to different network architectures, such as Transformer networks (Vaswani et al., 2017), where pre-training, using a large task related dataset, could also be explored.

Part V

Appendices

TSM Dataset Reference File Listings

A Training Subset

Table 1: Time-scale modification dataset reference file listing for the training subset.

File	Type	Description
Ardour_1.wav	Music	Saxophone and Rhythm Section
Brass_and_perc_5.wav	Music	Large Brass Ensemble
Brass_and_perc_6.wav	Music	Large Brass Ensemble
Brass_and_perc_10.wav	Music	Large Brass Ensemble
C_minor_Aeolian.wav	Music	Layered Synthesizers and Percussion
Chiptune.wav	Music	8-bit style Music
Dorothy_2.wav	Music	Music and Sound Effects
Electropop_6.wav	Music	Drums, Bass, Panning Electric Piano and Percussion

Continued on next page

Continued from previous page

File	Type	Description
G_Minor.wav	Music	Lead Synthesizer, Piano and Percussion
I_Believe_it_1.wav	Music	Synthesizers with delay effect
Jazz_1.wav	Music	Small Jazz Ensemble
Jazz_2.wav	Music	Small Jazz Ensemble
Oboe_cello_2.wav	Music	Classical Oboe and Cello
Oboe_piano_2.wav	Music	Classical Oboe and Piano
Oboe_piano_3.wav	Music	Classical Oboe and Piano
Oriental_2.wav	Music	Ambient Tuned and Untuned Percussion
Rock_2.wav	Music	Guitars, Bass and Drums
Saxophones_1.wav	Music	Classical Saxophone Quartet
Saxophones_3.wav	Music	Classical Saxophone Quartet
Synth_Pad_1.wav	Music	Slow evolving buzzy pad synthesizer
There_is_no_way.wav	Music	Sound Effects and Male Speech
Voice_and_Piano_4.wav	Music	Operatic singing with Piano
Voice_and_Piano_5.wav	Music	Operatic singing with Piano
Woodwinds_3.wav	Music	Classical Reed Woodwind Group
Woodwinds_5.wav	Music	Classical Reed Woodwind Group
Yellow_1.wav	Music	Classical Orchestra
Yellow_2.wav	Music	Violin, Flute and Triangle
Alto_Sax_06.wav	Solo	Dry Alto Saxophone
Alto_Sax_07.wav	Solo	Dry Alto Saxophone
Alto_Sax_08.wav	Solo	Dry Alto Saxophone
Alto_Sax_16.wav	Solo	Dry Alto Saxophone

Continued on next page

Continued from previous page

File	Type	Description
Alto_Sax_17.wav	Solo	Dry Alto Saxophone
Bass_2.wav	Solo	Tubby Synth Bass
Drums_2.wav	Solo	Simple Drum Kit Pattern
E_piano_1.wav	Solo	Electric Piano with wide auto-pan
Guitar.wav	Solo	Jazz chords
I_Believe_it_3.wav	Solo	Lead Synthesizer
Mexican_Flute_01.wav	Solo	Solo flute from Mexico, non-standard tuning
Music_Box_01.wav	Solo	Individual and chord from small music box
Ocarina_01.wav	Solo	Solo Ocarina
Perc_2.wav	Solo	Unpitched Percussion
Percussion.wav	Solo	Gong and Bass Drum
Piano_Allegro.wav	Solo	Solo Piano Recording
Piano_Turkish.wav	Solo	Solo Piano Recording
Rain_Stick_St_04.wav	Solo	Stereo Shaker
Rain_Stick_St_05.wav	Solo	Standard Rainstick sound
Saxophones_4.wav	Solo	Reverberant Solo Saxophone
Shaker_01.wav	Solo	Small light shaker
Solo_flute_1.wav	Solo	Reverberant Solo Flute
Solo_flute_3.wav	Solo	Reverberant Solo Flute
Sop_Sax_05.wav	Solo	Dry Soprano Saxophone
Synth_Bass_1.wav	Solo	Tight and Deep Synth Bass
Tambourine_01.wav	Solo	Single Tambourine Hits
Tambourine_02.wav	Solo	Shaking Tambourine

Continued on next page

Continued from previous page

File	Type	Description
Thumb_Piano_01.wav	Solo	Pitched Percussion
Triangle_01.wav	Solo	Rhythm played on Triangle
Vaccum_cleaner_2.wav	Solo	Descending chirp-like sound
Yellow_Violin.wav	Solo	Dry Solo Violin
Child_1.wav	Voice	Female "Once upon a time in a land far far away."
Child_2.wav	Voice	Female "The funniest clown that anyone had ever seen."
Child_3.wav	Voice	Female "Boo always managed to make every single child laugh."
Choral.wav	Voice	Reverberant large choir
Farmacist_2.wav	Voice	Comic Male "I've got a word for her."
Female_1.wav	Voice	"If my mouth did not permit me to speak."
Female_5.wav	Voice	"When you look into my eyes."
Female_6.wav	Voice	"Are you as fearful as I?"
Female_8.wav	Voice	"Alright, how do I run?"
Female_9.wav	Voice	"And yet you're still rubbish."
Female_10.wav	Voice	"Well it's not exactly what I want, but it's certainly what I need."
Female_11.wav	Voice	"For once I want a horrible meal I can complain about."
Female_12.wav	Voice	"The good experiences just become boring."

Continued on next page

Continued from previous page

File	Type	Description
Female_BV_1.wav	Voice	Three part harmony
Female_BV_2.wav	Voice	Three part harmony
Female_opera.wav	Voice	Solo female voice
Male_2.wav	Voice	"No he doesn't, come on let's go."
Male_5.wav	Voice	"10 seconds take your best shot."
Male_7.wav	Voice	"I have a family."
Male_8.wav	Voice	"What did you mean before when you said I, had a little girl?"
Male_9.wav	Voice	"I am sitting in a room."
Male_12.wav	Voice	"Do you want some more water Murray?"
Male_13.wav	Voice	"It was losing the horses that got us here."
Male_15.wav	Voice	"Told him I'd bring him back a Kangaroo skull."
Male_18.wav	Voice	"That you'd been looking for one of these for your boy."
Male_19.wav	Voice	"If it's what you need, then it's the best possible thing to do."
Male_21.wav	Voice	"Why do you even come to the pub if you're never going to drink anything?"
Male_23.wav	Voice	"I can pay for it if you like."
Male_24.wav	Voice	"I'm always going to have two cents that I can do nothing with."

Continued on next page

Continued from previous page

File	Type	Description
Male_sing_1.wav	Voice	"Words can't describe her."

B Testing Subset

Table 2: Time-scale modification dataset reference file listing for the testing subset.

File	Type	Description
Ardour_2.wav	Music	Saxophone and Rhythm Section
Brass_and_perc_9.wav	Music	Large Brass Ensemble
Jazz_3.wav	Music	Small Jazz Ensemble
Oboe_piano_1.wav	Music	Classical Oboe and Piano
Rock_4.wav	Music	Guitars, Bass and Drums
Saxophones_6.wav	Music	Classical Saxophone Quartet
Woodwinds_4.wav	Music	Classical Reed Woodwind Group
Alto_Sax_15.wav	Solo	Dry Alto Saxophone
Mexican_Flute_02.wav	Solo	Solo flute from Mexico, non-standard tuning
Ocarina_02.wav	Solo	Solo Ocarina
Solo_flute_2.wav	Solo	Reverberant Solo Flute
Synth_Bass_2.wav	Solo	Tight and Deep Synth Bass
Triangle_02.wav	Solo	Roll played on Triangle
Child_4.wav	Voice	Female “He never even tried.”
Female_2.wav	Voice	“Could my thoughts be expressed through my body alone?”
Female_4.wav	Voice	“Could it be protecting me against possible scrutiny?”
Male_6.wav	Voice	“Hey, just google it man.”
Male_16.wav	Voice	“I’m going down to the water.”

Continued on next page

Continued from previous page

File	Type	Description
Male_22.wav	Voice	"Sometimes you got to do what you want to do."
You_mean_this_one.wav	Voice	Sound Effects and Female Speech

C Unused Reference File Subset

Table 3: Time-scale modification dataset reference file listing for unused files

File	Type	Description
Brass_and_perc_3.wav	Music	Large Brass Ensemble
Brass_and_perc_4.wav	Music	Large Brass Ensemble
Brass_and_perc_7.wav	Music	Large Brass Ensemble
Brass_and_perc_8.wav	Music	Large Brass Ensemble
Dorothy_1.wav	Music	Music and Sound Effects
Electropop_1.wav	Music	Drums, Bass, Panning Electric Piano and Percussion
Electropop_3.wav	Music	Drums, Bass, Panning Electric Piano and Percussion
I_Believe_it_2.wav	Music	Synthesizers with delay effect
Oboe_cello_1.wav	Music	Classical Oboe and Cello
Oriental_3.wav	Music	Ambient Tuned, Untuned Percussion, Rainstick
Rock_1.wav	Music	Guitars, Bass and Drums
Saxophones_2.wav	Music	Classical Saxophone Quartet
Saxophones_5.wav	Music	Classical Saxophone Quartet
Voice_and_Piano_1.wav	Music	Operatic singing with Piano
Voice_and_Piano_2.wav	Music	Operatic singing with Piano
Woodwinds_1.wav	Music	Classical Reed Woodwind Group
Woodwinds_2.wav	Music	Classical Reed Woodwind Group
Yellow_3.wav	Music	Classical Orchestra
Alto_Sax_02.wav	Solo	Dry Alto Saxophone

Continued on next page

Continued from previous page

File	Type	Description
Alto_Sax_05.wav	Solo	Dry Alto Saxophone
Bass_4.wav	Solo	Tubby Synth Bass
E_piano_6.wav	Solo	Electric Piano with wide auto-pan
Perc_1.wav	Solo	Pitched Percussion
Rain_Stick_St_03.wav	Solo	Rhythmic Shaker
Sop_Sax_02.wav	Solo	Dry Soprano Saxophone
Sop_Sax_04.wav	Solo	Dry Soprano Saxophone
Thumb_Piano_02.wav	Solo	Pitched Percussion
Thunder_tube_01.wav	Solo	Thunder type sound effect
Vaccum_cleaner_1.wav	Solo	Descending chirp-like sound
Child_5.wav	Voice	Female "There was a hungry wolf lurking nearby."
Farmacist_1.wav	Voice	Comic Male "Irritatingly wordy, artificially intelligent, meticulously articulate."
Farmacist_3.wav	Voice	Comic Male "I just want to hurt her."
Female_3.wav	Voice	"Is everything ok?"
Female_7.wav	Voice	"Harden up princess."
Male_1.wav	Voice	"Well this has been fantastic, don't you have class?"
Male_3.wav	Voice	"You're not even any good at the game."
Male_4.wav	Voice	"At least I've got air support."
Male_10.wav	Voice	"Maybe no more than a day ago."

Continued on next page

Continued from previous page

File	Type	Description
Male_11.wav	Voice	"I can see her footprints in the sand there."
Male_14.wav	Voice	"Oh it's not as bad as it looks."
Male_17.wav	Voice	"Rests on your head as much as Murray's."
Male_20.wav	Voice	"Actually I'd love to hear it, I love it when things go wrong with you. It makes me feel better about myself."

Bibliography

A. Altoe. A transient-preserving audio time-stretching algorithm and a real-time realization for a commercial music product. Master's thesis, Faculty of Engineering, University of Padova, Padua, Italy, 12 2012.

C. Avendano and J. Jot. Frequency domain techniques for stereo to multichannel upmix. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pages 1–10, Espoo, Finland, 2002.

A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke. Non-intrusive speech quality assessment using neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635. IEEE, 2019.

G. Ballou. *Handbook for Sound Engineers*. Focal Press, Oxford, UK, 2008.

J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—Temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384, 2013.

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5):1035–1047, 2005.

J. Bertrand, P. Bertrand, and J. Ovarlez. Discrete mellin transform for signal analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1603–1606. IEEE, 1990.

J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997.

J. Bonada. Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *International Computer Music Conference*, 2000.

J. Bonada. *Audio Time-Scale Modification in the Context of Professional Audio Post-Production*. PhD thesis, Graduate Division, University of Pompeu Fabra, Barcelona, Spain, 2002.

A. A. Catellier and S. D. Voran. Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 331–335, 2020. doi: 10.1109/ICASSP40776.2020.9054204.

M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines. ViSQOL v3: An open source production ready objective speech and audio metric. *arXiv preprint arXiv:2004.09584*, 2020.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

E. Damskägg and V. Välimäki. Audio time stretching using fuzzy classification of spectral bins. *Applied Sciences*, 7(12):1293, 2017.

T. Dau, B. Kollmeier, and A. Kohlrausch. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, 102(5):2892–2905, 1997.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

D. Dorran, R. Lawlor, and E. Coyle. An efficient phasiness reduction technique for moderate audio time-scale modification. In *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx-04)*, Naples, Italy, 2004.

D. Dorran, R. Lawlor, and E. Coyle. Multi-channel audio time-scale modification. In *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

J. Driedger and M. Muller. TSM toolbox: MATLAB implementations of time-scale modification algorithms. In *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, pages 1–8, Erlangen, Germany, 2014.

J. Driedger and M. Muller. A review of time-scale modification of music signals. *Applied Sciences*, 6(2):57–83, 2016.

J. Driedger, M. Muller, and S. Ewert. Improving time-scale modification of music signals using harmonic-percussive separation. *IEEE Signal Processing Letters*, 21 (1):105–109, 2014.

J. Durrant and J. Lovrinic. *Bases of Hearing Science*. Williams & Wilkins, 1995. ISBN 9780683027372. URL https://books.google.com.au/books?id=7a1oQgAACAAJ.

V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2011.

T. H. Falk and W.-Y. Chan. Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1935–1947, 2006.

H. Fastl and E. Zwicker. *Psychoacoustics: facts and models*, volume 22. Springer Science & Business Media, 2006.

L. Fierro and V. Välimäki. Towards objective evaluation of audio time-scale modification methods. In *Proceedings of the 17th Sound and Music Computing Conference*, pages 457–462, 2020.

J. Flanagan and R. Golden. Phase vocoder. *Bell System Technical Journal*, 45(9): 1493–1509, 1966.

J. L. Flanagan. The ear and hearing. In *Speech Analysis Synthesis and Perception*, pages 86–140. Springer, 1972.

H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, 5(2):82–108, 1933. doi: 10.1121/1.1915637. URL https://doi.org/10.1121/1.1915637.

S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang. Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. In *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, 2018.

H. Gamper, C. K. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke. Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 85–89. IEEE, 2019.

godock. EAQUAL, 2017. URL https://github.com/godock/eaqual.

A. M. Gomez, B. Schwerin, and K. Paliwal. Objective intelligibility prediction of speech by combining correlation and distortion based techniques. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

G. Gottardi. Perceptual evaluation of audio quality, 2013. URL http://peaqb.sourceforge.net.

D. Griffin and J. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

M. Hansen and B. Kollmeier. Objective modeling of speech quality with a psychoacoustically validated auditory model. *Journal of the Audio Engineering Society*, 48 (5):395–409, 2000.

W. W. Hauck and S. Anderson. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, 12(1):83–91, 1984.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

M. Holters. GstPEAQ - a gstreamer plugin for perceptual evaluation of audio quality (PEAQ), 2017. URL https://github.com/HSU-ANT/gstpeaq.

Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1): 229–238, 2007.

X. Huang, A. Acero, H. Hon, and R. Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*, volume 95. Prentice hall PTR Upper Saddle River, 2001.

R. Huber and B. Kollmeier. PEMO-Q — A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on audio, speech, and language processing*, 14(6):1902–1911, 2006.

ircamLab. Ircam ts (version 1.0.11) [computer program], 2014. URL http://www.ircamlab.com/products/p1680-TS/. (Last viewed June 11, 2020).

ITU-T. ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Technical report, ITU, Tech. Rep, 1997.

ITU-T. ITU-R BS. 1387-1: Method for objective measurements of perceived audio quality. Technical report, ITU, Tech. Rep, 2001a.

ITU-T. ITU-R P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical report, ITU, Tech. Rep, 2001b.

ITU-T. ITU-R BS.1283-1: A guide to ITU-R recommendations for subjective assessment of sound quality. Technical report, ITU, Tech. Rep, 2003.

ITU-T. ITU-T P.10: Vocabulary for performance and quality of service, amendment 2: New definitions for inclusion in recommendation ITU-T P. 10/G. 100. Technical report, ITU, Tech. Rep, 2008.

ITU-T. ITU-R BS.1534-1 method for the subjective assessment of intermediate quality level of coding systems. Technical report, ITU, Tech. Rep, 2014.

ITU-T. ITU-R BS. 1284-1: General methods for the subjective assessment of sound quality. Technical report, ITU, Tech. Rep, 2019.

izotope. RX 7 variable time [computer program], 2018. URL https://www.izotope.com/en/products/rx/features/variable-time.html. (Last viewed June 12, 2020).

S. Jamieson et al. Likert scales: how to (ab)use them. *Medical education*, 38(12): 1217–1218, 2004.

N. Jillings, D. Moffat, B. De Man, and J. D. Reiss. Web Audio Evaluation Tool: A browser-based listening test environment. In *12th Sound and Music Computing Conference*, pages 147–152, Maynooth, Ireland, July 2015.

Z. Jun, T. Wei, C. Yanpu, and G. Yue. Parameters evaluation of sola algorithm for time scale modification. *International Journal of Speech Technology*, 10(2-3):89–94, 2007.

P. Kabal et al. An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality. *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, pages 1–89, 2002.

G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou. Time-scale modifications based on a full-band adaptive harmonic model. In *Proceedings of ICASSP '13*, pages 8193–8197. IEEE, 2013.

T. Karrer, E. Lee, and J. Borchers. PhaVoRIT: A phase vocoder for real-time interactive time-stretching. Technical report, Media Computing Group, 01 2006.

T. Kastner. Evaluating physical measures for predicting the perceived quality of blindly separated audio source signals. In *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

T. Kastner and J. Herre. An efficient model for estimating subjective quality of separated audio source signals. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 95–99. IEEE, 2019.

D.-S. Kim and A. Tarraf. ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality. *Bell Labs Technical Journal*, 12 (1):221–236, 2007.

D. Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1278–1281. IEEE, 1982.

D. Lakens. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

J. Laroche. *Time and Pitch Scale Modification of Audio Signals*, pages 279–309. Springer US, Boston, MA, 2002. doi: 10.1007/0-306-47042-X_7.

J. Laroche and M. Dolson. Phase-vocoder: About this phasiness business. In *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 1997.

J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013. ISBN 1466504218, 9781466504219.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2017.

L. Malfait, J. Berger, and M. Kastner. P. 563—the ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, 2006.

Melodyne. Melodyne (version 5) [computer program], 2020. URL https://www.celemony.com/en/melodyne. (Last viewed June 12, 2020).

A. Moinet and T. Dutoit. PVSOLA: A phase vocoder with synchronized overlap-add. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, pages 269–275, Paris, France, 2011.

E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.

E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech communication*, 16(2):175–205, 1995.

K. S. R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

A. Nicolson, J. Hanson, J. Lyons, and K. Paliwal. Spectral subband centroids for robust speaker identification using marginalization-based missing feature theory. *International Journal of Signal Processing Systems*, 6(1):12–16, 2018. ISSN 2315-4535. doi: 10.18178/ijsps.6.1.12-16.

N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, pages 139–144, 01 2008.

E. S. Ottosen and M. Dörfler. A phase vocoder based on nonstationary gabor frames. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11): 2199–2208, 2017.

B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley. AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech, 2016.

PaulStretch. Paul's extreme sound stretch (version 2.2-2) [computer program], 2011. URL http://hypermammut.sourceforge.net/paulstretch/. (Last viewed June 12, 2020).

J. O. Pickles. *An introduction to the physiology of hearing*, volume 2. Academic press London, 1988.

M. Portnoff. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech, And Signal Processing*, 24(3): 243–248, 1976.

M. Portnoff. Time-scale modification of speech based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):374–390, 1981.

J. Proakis and D. Manolakis. *Digital Signal Processing*. Pearson Prentice Hall, Upper Saddle River, N.J., 2007.

Z. Průša and N. Holighaus. Phase vocoder done right. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 976–980. IEEE, 2017.

M. Puckette. Phase-locked vocoder. In *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 222–225, New Paltz, N.Y., 1995. IEEE.

D. J. Putka, H. Le, R. A. McCloy, and T. Diaz. Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, 93(5):959, 2008.

T. F. Quatieri and R. J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, 1992.

E. Ravelli, M. Sandler, and J. P. Bello. Fast implementation for non-linear time-scaling of stereo signals. In *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx-05)*, pages 182–185, Madrid, Spain, 2005.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP'01).*, volume 2, pages 749–752. IEEE, 2001.

A. Röbel. A new approach to transient processing in the phase vocoder. In *6th International Conference on Digital Audtio Effects (DAFx)*, pages 298–305, 2003.

T. Roberts. A time-scale modification dataset with subjective quality labels, 2020. URL http://dx.doi.org/10.21227/ny9p-rv41.

T. Roberts and K. K. Paliwal. Stereo time-scale modification using sum and difference transformation. In *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–5. IEEE, 2018.

T. Roberts and K. K. Paliwal. Time-scale modification using fuzzy epoch-synchronous overlap-add (FESOLA). In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 31–34. IEEE, 2019.

T. Roberts and K. K. Paliwal. An objective measure of quality for time-scale modification of audio. *The Journal of the Acoustical Society of America*, -(-):–, 2020a. Under Review.

T. Roberts and K. K. Paliwal. A time-scale modification dataset with subjective quality labels. *The Journal of the Acoustical Society of America*, 148(1):201–210, 2020b. doi: 10.1121/10.0001567. URL https://doi.org/10.1121/10.0001567.

G. Roma, O. Green, and P. Tremblay. Time scale modification of audio using non-negative matrix factorization. In *Proc. of the 22nd Int. Conference on Digital Audio Effects (DAFx-19)*, pages 1–6, Birmingham, UK, 2019.

S. Roucos and A. Wilgus. High quality time-scale modification for speech. In *Proceedings of ICASSP '85*, volume 10, pages 493–496. IEEE, 1985.

S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula. Epoch-synchronous overlap-add (ESOLA) for time-and pitch-scale modification of speech signals. *arXiv preprint arXiv:1801.06492*, 2018. unpublished.

M. Ryan and M. Frater. *Communications and Information Systems*. Argos Press P/L, Yarralumla, AUS, 2002.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes. A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*, 80:84–94, 2016.

N. Sharma, S. Potadar, S. R. Chetupalli, and T. Sreenivas. Mel-scale sub-band modelling for perceptually improved time-scale modification of speech and audio signals. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–5. IEEE, 2017.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

A. Sorensen and H. Gardner. Systems level liveness with extempore. In *Proceedings of the 2017 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pages 214–228, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5530-8. doi: 10.1145/3133850.3133858. URL http://doi.acm.org/10.1145/3133850.3133858.

G. A. Soulodre and M. C. Lavoie. Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999.

S. S. Stevens. A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5):405, 1936.

R. C. Streijl, S. Winkler, and D. S. Hands. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22 (2):213–227, 2016.

B. Sylvestre and P. Kabal. Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation. In *Proceedings of ICASSP '92*, volume 1, pages 81–84. IEEE, 1992.

S. M. Thampi, A. Gelbukh, and J. Mukhopadhyay. *Advances in signal processing and intelligent recognition systems*. Springer, 2014.

T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes. PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

M. Torcoli. An improved measure of musical noise based on spectral kurtosis. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 90–94. IEEE, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of ICASSP '93*, volume 2, pages 554–557. IEEE, 1993.

E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

J. Volkmann, S. S. Stevens, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):208–208, 1937. doi: 10.1121/1.1901999. URL https://doi.org/10.1121/1.1901999.

S. Welch and M. Cohen. Perceptual coding in python, 2015. URL https://github.com/stephencwelch/Perceptual-Coding-In-Python.

F. M. Wiener and D. A. Ross. The pressure distribution in the auditory canal in a progressive sound field. *The Journal of the Acoustical Society of America*, 18 (2):401–408, 1946. doi: 10.1121/1.1916378. URL https://doi.org/10.1121/1.1916378.

B. Xiao and Y. Jiang. The comparison of window functions for different subbands in phase vocoder. *3rd International Conference on Multimedia Technology*, pages 1465–1472, 11 2013.

R. Yoneguchi and T. Murakami. Time-scale and pitch-scale modification by the phase vocoder without occurring the phase unwrapping problem. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017.

U. Zölzer, X. Amatriain, D. Arfib, J. Bonada, G. D. Poli, P. Dutilleux, et al. *DAFX - Digital Audio Effects*. John Wiley & Sons, 2002. ISBN 9780471490784. URL https://books.google.com.au/books?id=h90HIV0uwVsC.

Zplane Development. Èlastique time stretching & pitch shifting SDKs (version 3.2.5) [computer program], 2019. URL http://licensing.zplane.de/technology#elastique. (Last viewed October 31, 2019).

E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961. doi: 10.1121/1.1908630. URL https://doi.org/10.1121/1.1908630.

zynaptiq. ZTX time stretching and pitch shifting [computer program], 2015. URL https://www.zynaptiq.com/ztx/. (Last viewed June 28, 2020).