Doctoral Dissertation

Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

Bagus Tris Atmaja

Supervisor: Professor Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology

Information Science
July 17, 2020

Abstract

Humans perceive information in multimodal ways. Among many modalities, hearing is an important modality to perceive emotion from speech. Within speech, not only acoustic information can be extracted but also linguistic information. This linguistic information is commonly extracted via speech-to-text application. While the conventional paradigm in speech emotion recognition (SER) is performed by using acoustic information only, a new paradigm involves multimodal processing from multi-channel information. This research aims to propose methods for dimensional SER by combining acoustic and linguistic information. The problem thus exists on how to fuse both information. The following strategies are studied to solve this problem: SER by using acoustic features only, combining acoustic and linguistic information at the feature level, and combining acoustic and linguistic information at the decision level.

The first strategy aims to maximize the potency of recognizing dimensional emotion from acoustic information only. In this study, several acoustic features sets have been evaluated on both low-level and high-level features. Although high statistical functions might be limited in feature size, this kind of feature might be more informative than a local feature since it represents information within an utterance (by mean values) and captures the dynamic between frames (by standard deviation). This study reveals the effectiveness of means and standard deviations from a specific feature set for dimensional SER. Although several approaches has been carried out, acoustic-based SER has a limitation on low score of valence prediction.

A method to improve acoustic-based SER is by fusing acoustic and linguistic information. Linguistic information has been reported more predictive than acoustic information in predicting valence. Two fusing methods for acoustic-linguistic information fusion are studied: early-fusion approach and late-fusion approach. In the feature level (FL) early-fusion approach, two fusion methods are evaluated – feature concatenation and network concatenation. The FL methods show significant performance improvement over unimodal dimensional SER. In the second method using decision level (DL) late-fusion approach, acoustic and linguistic information are trained independently, and the results are fused by SVM to make the final predictions. Although this proposal is more complex than the previous FL fusion, the results show improvements over previous DL approach.

This research links the current problem in dimensional SER with its potential solution. The combination of acoustic and linguistic information fills the gap in dimensional SER. The FL approach improves the performance of unimodal SER significantly. The DL approach improves the FL approach's performance by mimicking human multimodal information fusion. The results devote insights for future strategy in implementing SER, whether to use acoustic-only features (less complex, less accurate), an early-fusion method (more complex, more accurate), or a late-fusion method (most complex, most accurate).
Keywords Dimensional emotion, speech emotion recognition, information fusion, affective computing.

Acknowledgments

The author wishes to express his sincere gratitude to his principal supervisor, Professor Masato Akagi of Japan Advanced Institute of Science and Technology, for his constant encouragement and kind guidance during this dissertation work. Prof. Akagi not only guides the author on the research direction but also kept the timeline of the research stay on the track. Prof. Akagi is a role model of supervisor for granting "license" for prospectus researcher, a PhD student.

The author also wishes to express his thanks to Professor Unoki, the co-supervisor of this dissertation. Without his critical questions, some part of this doctoral study will not exist. I owe the idea of multitask learning and silence pause features calculations to Prof. Unoki.

The author is grateful to Professor Kiyoaki Shiari for his helpful suggestions and discussions during minor research. With a background of acoustic information science, I have no sufficient knowledge of linguistic information processing until I conduct my minor research in his lab. His patience and kindness accelerate my understanding on the use of linguistic information for dimensional speech emotion recognition.

The kind, friendly, and warm atmosphere at Acoustic Information Science-Lab (Akagi and Unoki Lab) was the best place for conducting research and study. Surrounded by forests and natural landscape, there is no better place for my PhD study except in JAIST. I would like thank to all AIS members for their kindness support during my PhD studies.

Funding is one of crucial factor when conducting research. I would like to thank to Ministry of Education, Culture, Sports, Science and Technology (MEXT) for granting me a scholarship for my research and doctoral studies.

I would like to thank to my family and friends. This thesis is dedicated for them: my parents, my sisters, my wife and my son.

Finally, life is short, let's do the best.

Contents

Abstract i

Acknowledgments ii

1 Introduction 1
- 1.1 Research Aims . . . 1
- 1.2 Dissertation Contributions . . . 1
- 1.3 Dissertation Structure . . . 1

2 Literature Review 2
- 2.1 Introduction . . . 2
- 2.2 Emotion Models . . . 2
- 2.3 Datasets . . . 2
- 2.4 Features . . . 2
- 2.5 Classifiers . . . 2
  - 2.5.1 SVM . . . 2
  - 2.5.2 MLP . . . 2
  - 2.5.3 LSTM . . . 2
  - 2.5.4 CNN . . . 2
- 2.6 Summary . . . 2

3 Research Methodology 3
- 3.1 Research Philosophy . . . 3
- 3.2 Research Strategy . . . 4
- 3.3 Datasets . . . 4
- 3.4 Evaluation Metric . . . 4

4 Speech Emotion Recognition Using Acoustic Features 5
- 4.1 SER using Low-level Acoustic Features . . . 5
- 4.2 SER using High-level Acoustic Features . . . 9
- 4.3 Contribution of Silent Pause Features . . . 9
- 4.4 Acoustic Features Aggregation . . . 9
- 4.5 Summary . . . 9

5 Fusing Acoustic and Linguistic Information at Feature Level 10
- 5.1 Early fusion by features concatenation . . . 10
- 5.2 Early fusion by network concatenation . . . 10
- 5.3 Comparing ASR Outputs with Manual Transcription . . . 10

5.4 Summary 10

6 Fusing Acoustic and Linguistic Information at Decision Level 11
6.1 Two-stage Dimensional SER 11
6.1.1 LSTM network for unimodal prediction 11
6.1.2 SVR for results fusion 11
6.2 Results and Benchmarking 11
6.3 Summary 11

7 Conclusions 12
7.1 Summary 12
7.2 Future Research Directions 12

References 13

Publications 14

List of Figures

3.1 A concept used in this research: acoustic and linguistic information are extracted from (speech) data to obtain knowledge – the dimensional emotion. 3

4.1 Hann window and its spectrum 6
4.2 An example of Hamming window (middle) applied to sinusoid signal (left); the resulted windowed signal (right) is multiplication of both. 6
4.3 Frame-based processing for extracting low-level descriptors of an acoustic signal; the signal is an excerpt of IEMOCAP utterance with 400 samples frame length and 160 samples hop length; sampling frequency is 16 kHz. . 7

List of Tables

4.1 Acoustic feature sets: GeMAPS [1] and pyAudioAnalysis [2]. The numbers in parentheses indicate the total numbers of features (LLDs). 9

Chapter 1

Introduction

Speech emotion recognition (SER) is an emerging technologies resulted by science in multi discipline area including psychology, physiology, acoustics of physics, and affective computing. The later

1.1 Research Aims

1.2 Dissertation Contributions

1.3 Dissertation Structure

Chapter 2

Literature Review

2.1 Introduction

2.2 Emotion Models

2.3 Datasets

2.4 Features

2.5 Classifiers

2.5.1 SVM

2.5.2 MLP

2.5.3 LSTM

2.5.4 CNN

2.6 Summary

Chapter 3

Research Methodology

3.1 Research Philosophy

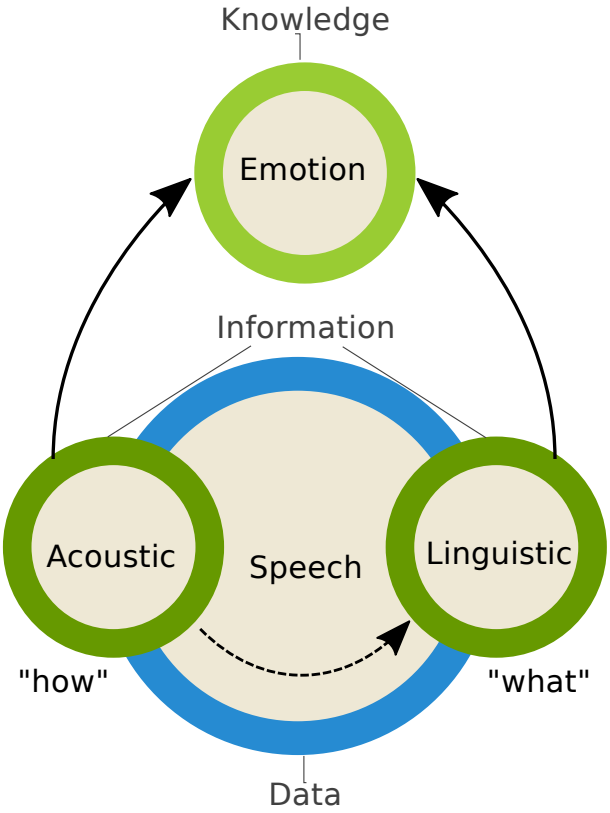

Figure 3.1: A concept used in this research: acoustic and linguistic information are extracted from (speech) data to obtain knowledge – the dimensional emotion.

3.2 Research Strategy

3.3 Datasets

3.4 Evaluation Metric

Chapter 4

Speech Emotion Recognition Using Acoustic Features

The purpose of this chapter is two-fold: (1) to investigate the effective region of analysis for acoustic feature extraction, whether frame-based region (local features) or utterance-based region (global features); and (2) to evaluate the effect of silence in dimensional speech emotion recognition. The latter issue is taken by investigating the effectiveness of removing silence vs. using silence as features. As auxiliary task, we performed an evaluation of aggregating acoustic features at input stage and compared the result with the baseline which is performed by aggregating output labels.

4.1 SER using Low-level Acoustic Features

SER in conventional ways are performed by extracting acoustic features on frame-based processing and then applied these features to a classifier. Let $y(n)$, with $n = 1, 2, 3, \ldots, L$, denotes acoustic signal with length L. In frame-based processing, this $y(n)$ signal is divided into frames by fixed length. A typical length for a single frame is 16-25 milliseconds (ms) with 10 ms to 15 ms hop length (stride). For 25 ms frame length and 10 ms hop length, which is equal to 60% overlap (15 ms), a window is applied to this frame to make the short-time signal behave as quasistationary signal – near stationary signal. In their original length, an acoustic signal vary with the time: non-stationary property. Windowing processes acoustic signal in short-term interval to remove this property. Figure 4.3 shows the windowing process; short-term signals seems more stationary than the original signal.

Windowing multiplies spectrum of input signal with window signal $w(n)$. A typical window function for acoustic signal is Hann and Hamming windows (named after Julius von Hann and Richard W. Hamming). The others are Rectangular, Bartlet, Kaiser, and Blackman. The choice of window function is based on two aspects: width of main lobe and additional lobes. Hann and Hamming windows only differ in weighting factors with similar concept: cosine-sum windows.

$$w[n] = A + B\cos\left(\frac{2\pi n}{M}\right), \qquad n = -M/2, \ldots, M/2 \tag{4.1}$$

where A is 0.5 for Hann and 0.54 for Hamming. B is 0.5 for Hann and 0.46 for Hamming. Both window functions are widely used in speech processing due to good trade-off between

time and frequency resolution (effect of side lobes). Figure 4.1 shows Hann window and its spectrum, while Figure 4.2 shows an example of Hamming window applied to sinusoid signal and its result.

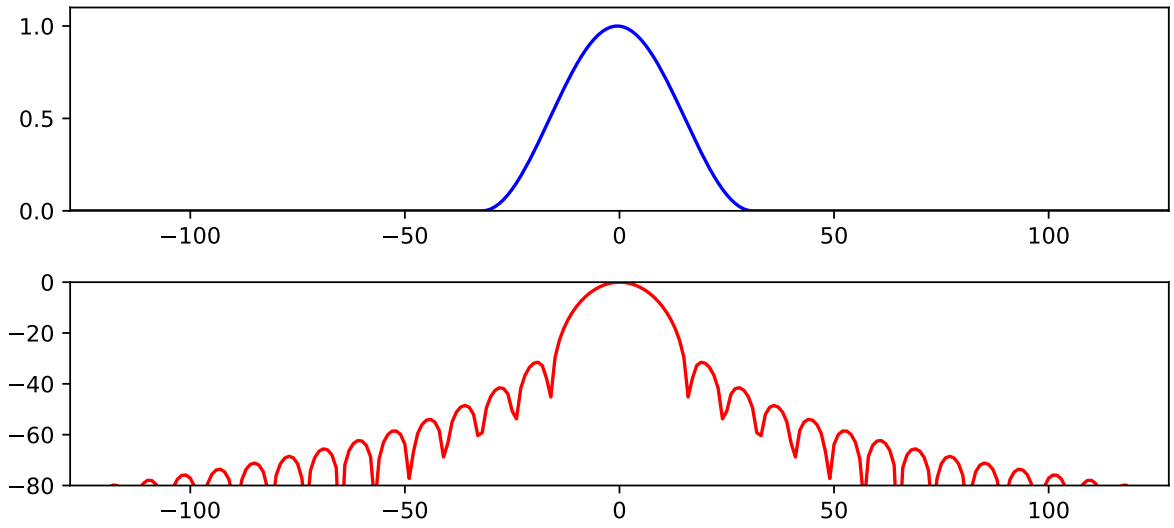

Figure 4.1: Hann window and its spectrum

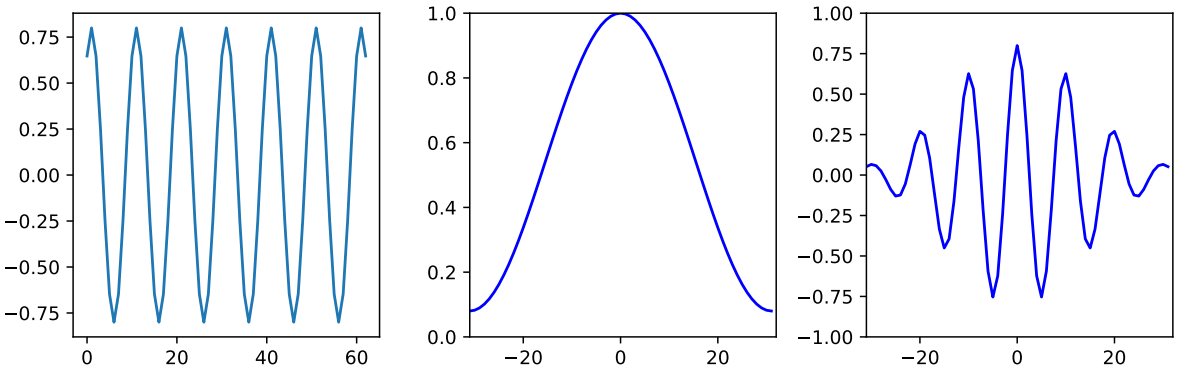

Figure 4.2: An example of Hamming window (middle) applied to sinusoid signal (left); the resulted windowed signal (right) is multiplication of both.

The length of a window is usually equal to the length of frame: one window per frame. If the length of a window is smaller than a frame, each frame will be windowed with window length and padded with zeros to match length of frame. It costs more computation. In speech emotion recognition, short window is used to capture short dynamics context while longer window is used to capture mid and longer dynamics. A common approach used short window to extract acoustic features in short-term time while long-term dynamics are modeled by statistical functions. Figure 4.3 shows frame-based processing of an acoustic signal (speech) which windows short-term signals using Hamming window.

The acoustic features extracted on each frame are known as local features or low-level descriptor (LLD) [3]. The most common LLD for speech processing is mel-frequency cepstral coefficients (MFCC). MFCC captures different aspects of the spectral shape of a

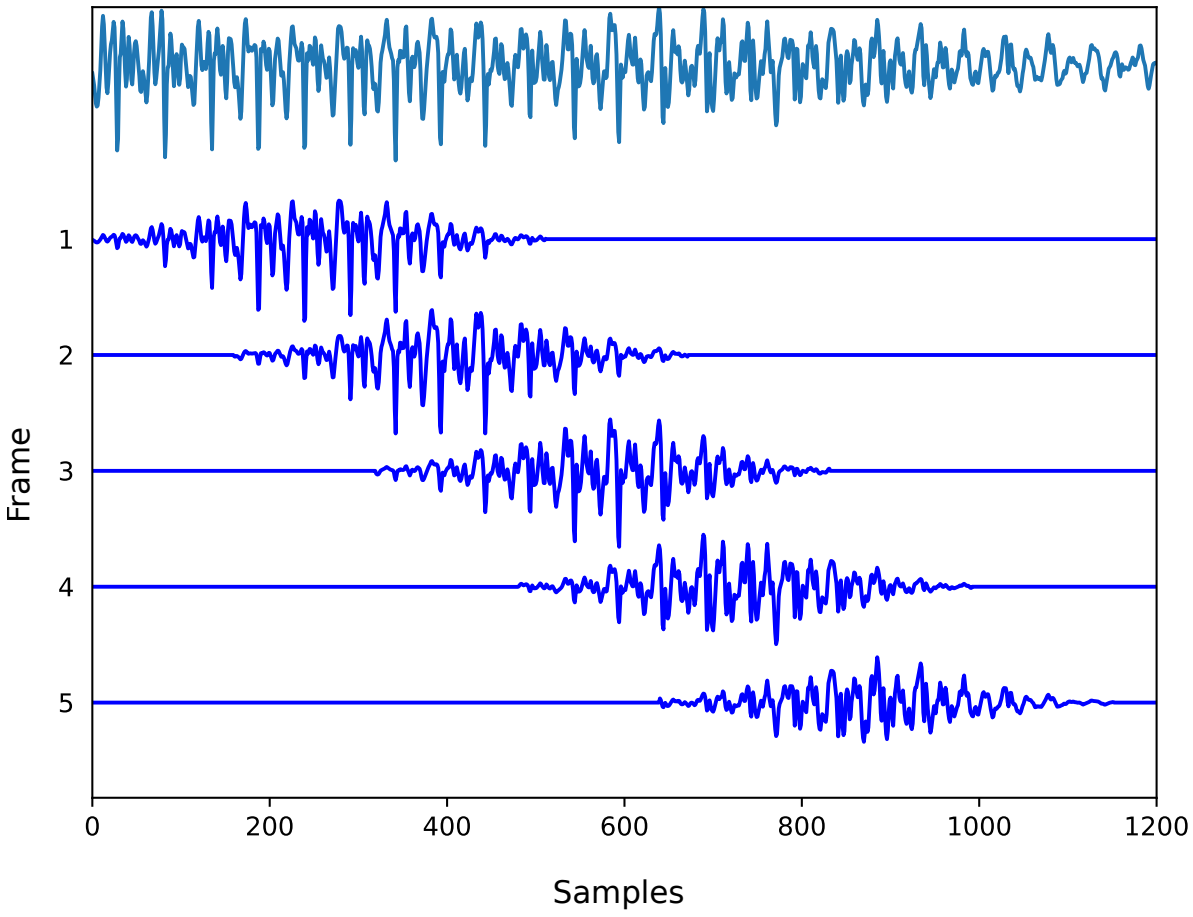

Figure 4.3: Frame-based processing for extracting low-level descriptors of an acoustic signal; the signal is an excerpt of IEMOCAP utterance with 400 samples frame length and 160 samples hop length; sampling frequency is 16 kHz.

speech. The following steps compute MFCC in sequences. First, FFT/DFT transformed time domain signal into frequency domain signal (spectra). Second, mel frequency warping function convert spectra in linear scale into mel scale. Although several functions have been proposed, a common approach keeps linear scale for acoustic frequencies below 1 kHz and converts to logarithmic scale for acoustic frequencies above 1 kHz. This conversion imitates human perceptual system. Third, convert a power spectrogram (amplitude squared) to decibel (dB) units (log). Finally, DCT computes MFCC as amplitude cepstra.

One of the important parameters in MFCC is the number of coefficients. A number of 13 to 40 coefficients are common for speech processing. For each frame 13 MFCCs are extracted. If there are 40 frames in an utterance, the dimension of MFCC features will be (40, 13). Obviously, the number of frames corresponds to the number of samples divided by hop length (in samples). If an utterance comprises 1 second (s) with 16000 Hz sampling rate, the number of samples is 16000. Using 25 ms (400 samples) window/frame length and 10 ms (160 samples) hop length, the number of frames is 16000/160, i.e., 100 frames. Figure **??** shows an MFCC spectrogram of an utterance with 40 coefficients.

Recently, researchers found that mel-spectrogram, also called as (mel) filter bank and mel frequency spectral coefficients (MFSC) for deep learning-based automatic speech recognition (ASR) (e.g., [4]). Given that deep learning system is less susceptible with high correlated input, the DCT step in previous MFCC calculation is not necessary since it is linear transformation. DCT discards some information in speech signals which are highly non-linear [5]. Furthermore, a logarithmic version of mel-spectrogram, i.e., log mel-spectrogram, is preferable one since deep learning learns better in this scale. The mel-spectrogram visualization as shown in Figure **??** support this argument in comparison with MFCC visualization.

Apart from the use of one type acoustic features for speech processing, some researchers have proposed a set of acoustic features for speech emotion recognition. Eyben et al. [1] proposed Geneva minimalistic parameter set (GeMAPS) as standard acoustic features for affective voice research. The proposed acoustic features are based on: (1) physiological changes in voice production, (2) proven significance in previous studies, and (3) theoretical significance. The proposed acoustic features comprises 23 LLDs as shown in Table 4.1. This acoustic feature set is extracted on frame-processing basis with 25 ms frame length and 10 ms hop length.

Giannakopulos [2] proposed pyAudioanalysis as an open source Python library for audio signal analysis. The library supports a wide range of audio analysis procedures such as feature extraction, classification, supervised and unsupervised segmentation, and visualization. Different from GeMAPS feature set, pyAudioanalysis targets a wide range of voice application like audio event detection, speech emotion recognition, music segmentation, and health application. The short-term feature set, which is extracted on frame-processing basis, consists of 34 LLDs. These LLDs are shown in Table 4.1.

Table 4.1: Acoustic feature sets: GeMAPS [1] and pyAudioAnalysis [2]. The numbers in parentheses indicate the total numbers of features (LLDs).

GeMAPs (23)	pyAudioAnalysis (34)
intensity, alpha ratio, Hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, harmonics-to-noise ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.	zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation.

4.2 SER using High-level Acoustic Features

4.3 Contribution of Silent Pause Features

4.4 Acoustic Features Aggregation

4.5 Summary

Chapter 5

Fusing Acoustic and Linguistic Information at Feature Level

This paper evaluates the error correcting capabilities of concatenated codes employing MDS codes as outer codes and time-varying randomly selected inner codes, used on discrete memoryless channels with modified MMI decoding. It is proved that Gallager's random coding error exponent can be obtained for all rates by such codes.

5.1 Early fusion by features concatenation

5.2 Early fusion by network concatenation

5.3 Comparing ASR Outputs with Manual Transcription

5.4 Summary

Chapter 6

Fusing Acoustic and Linguistic Information at Decision Level

This paper evaluates the error correcting capabilities of concatenated codes employing MDS codes as outer codes and time-varying randomly selected inner codes, used on discrete memoryless channels with modified MMI decoding. It is proved that Gallager's random coding error exponent can be obtained for all rates by such codes.

6.1 Two-stage Dimensional SER

6.1.1 LSTM network for unimodal prediction

6.1.2 SVR for results fusion

6.2 Results and Benchmarking

6.3 Summary

Chapter 7

Conclusions

This paper evaluates the error correcting capabilities of concatenated codes employing MDS codes as outer codes and time-varying randomly selected inner codes, used on discrete memoryless channels with modified MMI decoding. It is proved that Gallager's random coding error exponent can be obtained for all rates by such codes.

7.1 Summary

7.2 Future Research Directions

References

[1] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[2] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis/

[3] P. Herrera, X. Serra, and G. Peeters, "Audio Descriptors and Descriptor Schemes in the Context of MPEG-7," in *Int. Comput. Music Conf.*, 1999, pp. 581–584. [Online]. Available: http://mtg.upf.edu/files/publications/icmc99-perfe.pdf

[4] A. Mohamed, "Deep Neural Network acoustic models for ASR," Ph.D. dissertation, University of Toronto, 2014.

[5] H. M. Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," 2016. [Online]. Available: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Publications

[1] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," APSIPA Trans. Signal Inf. Process., vol. 9, no. May, p. e17, May 2020.

[2] B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4482–4486.

[3] B. T. Atmaja and M. Akagi, "The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in 10th International Conference on Speech Prosody 2020, 2020, no. May, pp. 26–30.

[4] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 519–523.

[5] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 40–44.