Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

by

Bagus Tris Atmaja

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato Akagi

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

Information Science
July 16, 2020

Abstract

This paper investigates the error correcting capabilities of concatenated codes employing maximum distance separable (MDS) codes as outer codes and time-varying randomly selected constant composition inner codes, used on discrete memoryless channels with modified maximum mutual information decoding. It is proved that such code can achieve Gallager's random coding error exponent for all rates, while both encoding and decoding of the codes do not depend on the channel.

Acknowledgments

The author wishes to express his sincere gratitude to his principal advisor Associate Professor Tomohiko Uyematsu of Japan Advenced Institute of Science and Techology for his constant encouragement and kind guidance during this work. The author would like to thank his advisor Professor Eiji Okamoto of Japan Advanced Institute of Science and Technology for his helpful discussions and suggestions.

The author also wishes to express his thanks to Associate Professor Kunihiko Hiraishi and Mineo Kaneko of Japan Advanced Institute of Science and Technology for their suggestions and continuous encouragements.

The author is grateful to Professor Yoji Kajitani of Tokyo Institute of Technology for their helpful suggestions and discussions.

The author is grateful to all who have affected or suggested his areas of research. Professor Te Sun Han of Electro-communication university inspired the author through his constant activities on information theory, and gave the author valuable suggestions and kind encouragements.

The author devotes his sincere thanks and appreciation to all of them, and his colleagues.

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 Concept 3

3.3 Research Strategy 3

3.4 Datasets 3

3.5 Evaluation Metric 3

4 Speech Emotion Recognition Using Acoustic Features **4**

4.1 SER using Low-level Acoustic Features 4

4.2 SER using High-level Acoustic Features 8

4.3 Contribution of Silent Pause Features 8

4.4 Acoustic Features Aggregation 8

4.5 Summary 8

References **9**

Publications **10**

List of Figures

4.1 Hann window and its spectrum 5
4.2 An example of Hamming window (middle) applied to sinusoid signal (left); the resulted windowed signal (right) is multiplication of both. 5
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; samling frequency is 16 kHz. . . 6

List of Tables

4.1 Acoustic feature sets: GeMAPS [1] and pyAudioAnalysis [2]. The numbers in parentheses indicate the total numbers of features (LLDs). 8

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

The problem of specifying the structure of codes which achieve Gallager's error exponent remains as a major problem in information theory and coding theory. On this subject, Thommesen[?] investigated the concatenated codes with maximum distance separable (MDS) outer codes and time-varying inner codes used on DMC with maximum likelihood decoding. He has proved that Gallager's error exponents are asymptotically obtained for all rates by such codes. On the other hand, Ahlswede and Dueck[?] investigated the code called "permutation code" obtained by permutating coordinates of a single codeword. They have proved that Gallager's error exponents are asymptotically obtained for all rates less than capacity, by employing maximum mutual information (MMI) decoding. Since both encoding and decoding of their codes do not depend on particular channel, and their codes yield the best asymptotic performance, they are usually called as "universal codes". However, it is not known whether universal codes can be obtained by other structures of codes.

This paper extends the results of Thommesen, and shows that there exist good universal codes in a class of concatenated codes. We investigate the decoding error probability of concatenated codes employing MDS codes as outer codes and time-varying randomly selected constant composition codes as inner codes, used on arbitrary DMC with modified MMI decoding. It is shown that Gallager's error exponents are universally obtained for all rates by the proposed codes, provided that the length of the code is sufficiently large.

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

The outer codes to be considered are maximum distance separable (MDS) codes. Let Γ denote an MDS code over $GF(q)$, where q is a prime power. Let K denotes its dimension and N its block length. Then the minimum distance D of Γ is given by $D = N - K + 1$. When we refer to the rate τ of the outer code, we mean the dimensionless rate $\tau = K/N$.

The symbols of the codewords in Γ are encoded into sequences in T_P^n by inner encoders, where P is a specified type of sequences. Suppose that all inner encoders have the same code length n, the inner encoders $g_1, \cdots, g_N$ are defined by the mappings

$$g_i : GF(q) \to T_P^n \qquad (1 \leq i \leq N),$$

where the inner encoders are not necessary to be one to one. The rate r of the inner encoders is defined by $r = \ln q/n$ (nats/symbol).

In what follows, the inner codes are selected randomly from the specified ensemble. Especially, we deal with the following P-ensemble as an ensemble of inner codes.

Definition 1 (P-ensemble) For a given outer code, a given type $P \in \mathcal{P}_n$, and a given block length n of the inner code, we select Nq sequences $g_i(u) \in T_P^n$, where $1 \leq i \leq N$ and $u \in GF(q)$, independently and uniformly over T_P^n.

3.1 Research Philosophy

3.2 Research Concept

3.3 Research Strategy

3.4 Datasets

3.5 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-base 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. 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.

Windowing multiplies spectrum of input signal with window signal $w(n)$. A typical window function for acoustic signal is Hann or 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 usually is 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 speech. The following steps compute MFCC in sequences. First, FFT/DFT transformed

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; samling frequency is 16 kHz.

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 is 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 of 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

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.-r. Mohamed, "Deep Neural Network acoustic models for ASR," p. 129, 2014. [Online]. Available: https://tspace.library.utoronto.ca/bitstream/1807/44123/1/Mohamed{_}Abdel-rahman{_}201406{_}PhD{_}thesis.pdf

[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] T. Uyematsu, K. Kikuchi and K. Sakaniwa: "Trellis Coded Modulation for Multilevel Photon Communication System," Proc. of Inter. Symp. on Inform. Theory and Its Applications '92, pp.582-587 (Nov. 1992).

[2] T. Uyematsu: "Efficient Maximum Likelihood Decoding Algorithms for Linear Codes over Z-Channel," IEICE Trans. Fundamentals, vol.E76-A, no.9, pp.1430-1436 (Sep. 1994).

[3] E. Okamoto, T. Uyematsu, M. Mambo: "Permutation Cipher Scheme Using Polynomials over a Field," IEICE Trans. on Information and Systems, E78-D, no.2, pp.138-142 (Feb. 1995).

[4] T. Uyematsu and E. Okamoto: "A Construction of Codes with Exponential Error Bounds on Arbitrary Discrete Memoryless Channels," submitted to IEEE Trans. on Information Theory.