Multiple Emotion Recognition from a Speech Utterance using Deep Neural Architectures

Botsa Kishore Kumar

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad-500032, India
kishore.botsa@research.iiit.ac.in

1. Motivation

Speech is the most simplest way of communication among the humans. This is an inspiration to carry a lot of research in this area. In spite of tremendous research in emotional speech processing, the ability of the algorithms is poor since they are unable to model different emotions of the speaker. Emotion processing from speech signal is a challenging research domain. Speech emotion recognition systems have many practical applications: in call centers to scrutinizing the state of the attendants while responding to their customers will make better service quality. They are used in on-board driving system, which collects the data about the emotional state of the driver to prevent accidents.

Index Terms: Emotion classification, Harmonic analysis, Excitation source features, Convolutional neural networks.

2. Key Issues Identified

Designing speech based emotion recognition system [1] is challenging due to the following reasons: (1) Finding the features that correlate well with the emotions [2, 3]. (2) Emotions are speaker and the environment dependent [4, 5]. (3) Transient emotions are ineffective when a speaker is in one particular emotion for a longer time.

Extraction of features which will differentiate different emotions is one of the major issues in designing an emotion recognition system. There are four groups of speech features. (1) local and global features, (2) qualitative features, (3) spectral features, and (4) TEO (Teager energy operator) based features. In the literature:

- It is observed that continuous (local and global) speech features such as F_0 and duration features carry emotion-related information [6–8].
- Some experiments done on voice quality features have drawn a relation among the detected emotion and the quality of the voice.
- Spectral features such as MFCC are extracted from the short-time representations of speech signal.
- TEO structure of the pitch contour is used as a feature to classify the different emotions in speech.

There is no specific feature which extract all the information. Therefore, feature extraction for effective emotion recognition is still a challenging problem. In the past few years, different classification models such as Support Vector Machines (SVM), decision trees, K-nearest Neighbors (KNN), and Neural Networks (NN). Deep learning is one of the most popular area of research in machine learning. The ability of CNN to learn high dimensional features is useful in developing speech-based emotion recognition systems.

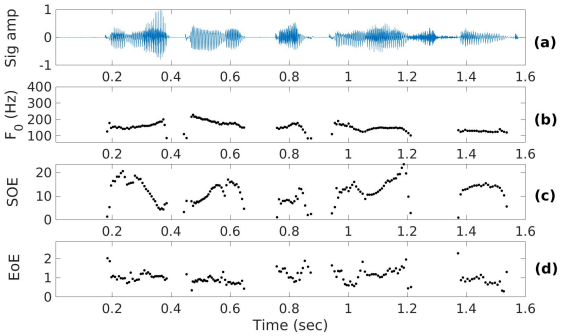

Figure 1: *Some Source/Prosody Features for "Neutral" Emotion (a) Speech Signal, (b) F_0 in Hz, (c) SoE, and (d) EoE.*

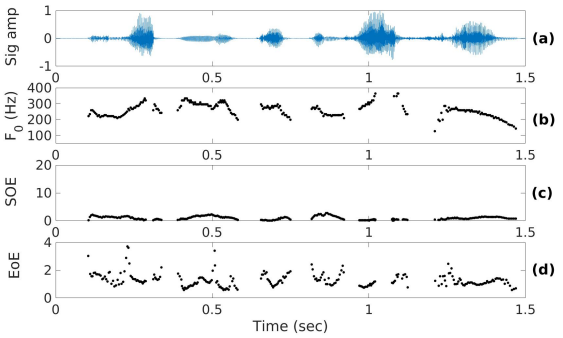

Figure 2: *Some Source/Prosody Features for "Anger" Emotion (a) Speech Signal, (b) F_0 in Hz, (c) SoE, and (d) EoE.*

3. Major Contributions

Speech production is a dynamic system which includes the vocal tract system and the excitation source. Each sound unit produced will have different duration and energy contours. Zero frequency filtering (ZFF) and linear prediction (LP) analysis are used to calculate the features. LP analysis is used to calculate linear prediction coefficients (LPCs) which represents the dynamically varying vocal tract system. Harmonic features are computed using adaptive Quasi-Harmonic Model (aQHM).

Harmonic representation gives an estimation of amplitudes and frequencies present in the speech signal. DFT based harmonic model gives sinusoidal frequencies and their corresponding amplitudes. Frequencies are computed by picking peaks from DFT spectrum. In DFT based method, window size is crucial in estimating parameters and resynthesizing of the signal.

Table 1: *Accuracy and Unweighted Average Recall (UAR) for the development set of EmotAsS database with different feature combination and classifiers (SVM, KNN, CNN).*

Features	Classifiers						Baseline System	
	SVM		KNN		CNN			
	Accuracy(%)	UAR(%)	Accuracy(%)	UAR(%)	Accuracy(%)	UAR(%)	Accuracy(%)	UAR(%)
S	50.8	32.46	54.82	26.46	51.6	41.9	45.3	37.8
S+H	47.2	31.7	48.4	25.7	47.4	40.27		

Table 2: *Results of test data using the CNN model trained with source and prosody features. Trail-1 corresponds to CNN with 7-Conv & 2-Pool and trail-2 corresponds to CNN with 6-Conv & 2-Pool.*

Model	Accuracy (%)	UAR (%)
Trail-1	40.013	32.758
Trail-2	45.111	29.298

Three models are taken into consideration for the study. In the analysis, a 2.5 ms frame rate is considered. The window size is fixed at 30 ms. Frame rate and analysis window size are same for both voiced and unvoiced regions. For each analysis frame, 10 frequencies corresponding to the highest amplitudes are computed.

There are many variants of the standard neural network. Convolutional Neural Network (CNN) is one among these variants. CNN possess a different network structure compared to deep neural networks (DNNs). It consists of two layers called convolution layer and pooling layer. Convolutional layer carry forward the inputs through some hidden units and fed to pooling layer which will allay the variability present in hidden units, which are due to different styles of speaking, channel disturbances. Speech signals will have these type of spectral variations. Hence, CNN are relatively more useful for speech processing than Deep Neural Networks (DNNs).

In my current research, harmonic features, excitation source features, and prosody features are used to classify different emotions. Harmonic amplitude variations are calculated using adaptive Quasi-Harmonic Analysis (QHA). The excitation source and prosody feature set consists of fundamental frequency (F0), Strength of Excitation (SoE), Energy of Excitation (EoE), duration features and their respective statistics. Convolutional Neural Network (CNN) is explored for the classification. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical Affect Sub-Challenge database (EmotAsS) is used for the study.

Some source/prosody features are shown in Figure 1 for the speech signal corresponding to emotion $Neutral$ and Figure 2 shows the speech signal corresponding to emotion $Anger$. In Figure 1 and 2 (a) represents speech signal corresponds to emotions, (b) represents fundamental frequency (F_0), (c) represents the strength of excitation (SoE), and (d) represents the energy of excitation ($E0E$). It is clear that F_0 of $Anger$ (in the range of 160 Hz to 370 Hz) is more than $Neutral$ (in the range of 100 Hz to 220 Hz). SoE is higher for $Neutral$ emotion (in the range of 0 dB to 20 dB) compared $Anger$ emotion (in the range of 0 dB to 3 dB). EoE is some what higher for anger compared to other emotions.

4. Results and Discussion

The performance of emotion classification system for different classifiers is given in Table 2. Accuracy and unweighted average recall (UAR) are used as performance metrics.

$$Accuracy = TP/(TP + FP) \quad (1)$$

$$UAR = TP/(TP + FN) \quad (2)$$

In addition to CNN, classification is done using Support Vector Machine (SVM), K-Nearest Neighbour (KNN). The results of CNN are better than other classification models and crossed the performance of baseline system. Baseline system is implemented using fusion based model trained with ComParE features. The model which gave best results on test data is used as baseline system and gave accuracy 45.3% and UAR 37.8% on development data

We have used the CNN model built using source and prosody features for evaluating the test data. The results are given in Table 2.

In trial-1, the test data applied to a CNN model trained with source and prosody features. The CNN model consists of seven convolution layers and two pooling layers. This model gave the accuracy 51.6%, UAR 41.9%. The $F1score$ of this model is found to be 27.26%. In tria-2, we used different CNN model with six convolution layers and two pooling layers which accuracy 48.04%, UAR 42.7%. The $F1score$ for trial-2 is found to be 25.72%. Though the UAR is better in trial-2, the performance is not better on test data considering UAR metric. This is because the $F1score$ is better in trial-1.

5. Future Plan

Each sentence may be uttered with different emotions with variable number like happy-laughter and sarcastic-laughter. The future plan of my research is to detect multiple emotions from the utterance using different deep classification models. The main concentration will be on hierarchical deep neural network models.

6. References

[1] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Canada., May 2004, pp. 577–580.

[2] Banse and Scherer, "Acoustic profiles in vocal emotion expression," *J Pers Soc Psychol*, vol. 1, pp. 614–636, March 1996.

[3] J. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[4] V. Hozjan and Z. Kacic, "Context-independent multilingual emotion recognition from speech signals," *International Journal of Speech Technology*, vol. 6, pp. 311–320, 2003.

[5] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[6] L. T. Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, no. 1, pp. 213–225, 2003.

[7] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.