SPEAKER DEPENDENCY OF SPECTRAL FEATURES AND SPEECH PRODUCTION CUES FOR AUTOMATIC EMOTION CLASSIFICATION

Vidhyasaharan Sethu[1, 2], Eliathamby Ambikairajah[1, 2] and Julien Epps[1]

[1]The School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052, Australia
[2]National Information Communication Technology Australia (NICTA),
Australian Technology Park, Eveleigh 1430, Australia

ABSTRACT

Spectral and excitation features, commonly used in automatic emotion classification systems, parameterise different aspects of the speech signal. This paper groups these features as speech production cues, broad spectral measures and detailed spectral measures and looks at how they differ in their performance in both speaker dependent and speaker independent systems. The extent of speaker normalisation on these features is also considered. Combinations of different features are then compared in terms of classification accuracies. Evaluations were conducted on the LDC emotional speech corpus for a five-class problem. Results indicate that MFCCs are very discriminative but suffer from speaker variability. Further, results suggest that the best front end for a speaker independent system is a combination of pitch, energy and formant information.

Index Terms— Emotion Classification, Feature comparison, MFCC, Group Delay, Gaussian mixture models

1. INTRODUCTION

Salovey *et al*. [1] defined emotional intelligence as having four branches: perception of emotion, emotions facilitating thought, understanding emotions and managing emotions. The lack of this emotional intelligence is one of the most significant differences between speech based human-human interaction and human-machine interaction. The focus of this paper is on the first of the four branches of emotional intelligence; namely, a system that is able to automatically detect the emotional state of a person based on speech.

This paper looks at an emotion classification system that does not utilize semantic or linguistic information. Such systems do not require any language models, and rely solely on prosodic and/or spectral features. Based on these features, classifiers such as neural networks [2], hidden Markov models (HMM) [3][4], Gaussian mixture models (GMM) [5] and support vector machines (SVM) [5] may be used to detect the emotional state of the speaker. A wide range of prosodic and spectral features have been proposed over the years for such systems [2-8]. Comparisons of the performances of commonly used features are available in [6-8]. The approach taken in these works are similar in that they adhere to a static modelling approach whereby frame level parameters such as pitch, energy, etc. are estimated and their statistics such as maximum, minimum, range, standard deviation, etc. are computed for each utterance to be used as features to a classification system. Consequently the comparisons are between different feature statistics (e.g. pitch range and standard deviation of energy) rather than speech parameters (e.g. pitch and energy).

An alternative to the static modelling approach is the dynamic modelling one where the frame level parameters are used as features directly. Huang *et al.* [4] suggest that Gaussian mixture models, when used in a dynamic modelling framework are able to model statistics like mean, range and standard deviation and they need not be computed explicitly. The results included in [4] and preliminary work in our labs supports this line of reasoning. A dynamic modelling approach would then allow for a comparison between speech parameters as opposed to a comparison between statistics and is the approach taken in the work reported in this paper.

Different features may have different levels of speaker dependent and emotion dependent characteristics. Consequently, they have differing performances in speaker dependent (trained on data from target speaker) and speaker independent (training and testing data come from different speakers) systems. Also, in some cases the information contained in a particular feature set could be complementary to the information in another set. This paper attempts to compare such features and determine if some or any of them are complementary.

2. SPECTRAL AND EXCITATION FEATURES

Features that have been reported to perform well in emotion classification tasks were selected to be compared to each other and are listed in this section. Since the system being studied does not make use of semantic or linguistic information, only acoustic, prosodic and spectral features were selected. Moreover, some popular features like speech rate do not fit into a dynamic modelling framework and hence were not considered.

Based on whether the features describe the speech spectrum or speech production parameters, they are classified as spectral features or speech production cues in this paper. The spectral features are further classified into broad and detailed spectral measures based on the level of spectral detail contained in them. While it has been shown that speaker variability in features significantly lowers the performance of a speaker independent system [9]; different features capture different amounts of the speaker's characteristics and consequently not all of them are affected to the same degree.

2.1. Speech Production Cues

Pitch and energy – These are the two most commonly used features in emotion classification. Pitch characterises the glottal excitation rate and energy is an estimate of the intensity of glottal excitation. Together, they characterise the glottal excitation in the standard speech production model and are used as a feature set. In this work, the YIN estimator [9] was used to estimate pitch and energy was computed as the mean squared value of the signal within each frame. Both single dimensional features were computed within frames of 40 ms duration (minimum duration for reliable pitch estimate) obtained using a rectangular window, with consecutive frames overlapping by 30ms.

Formants - The glottal excitation is spectrally shaped by the vocal tract in order to produce speech. The standard model of speech production models the vocal tract as an all pole filter whose resonances are termed formants. Of particular significance to voiced speech are the first three formants, which are characteristic of the sound produced. The first three formant frequencies and the corresponding formant energies, determined from the LPC magnitude spectrum, are concatenated to produce a 6 dimensional vector to characterise the vocal tract.

2.2. Detailed Spectral Measures (DSM)

Typical features in almost all speech processing characterise spectral information. Features that characterise spectral information in some detail are high dimensional features when compared to broad spectral measures.

Mel frequency cepstral coefficients – the MFCCs, which characterise the magnitude spectrum, are commonly used in speech processing, particularly in speech recognition and speaker recognition systems. In all experiments described herein, 12 dimensional MFCC vectors were used.

LPC Based Group Delay – The recently proposed LPC based group delay features are based on the all-pole filter model of speech production. While related to the spectral envelope of the signal, we have previously shown that the group delay features explicitly model formant bandwidths. This makes these features suitable for emotion classification and has been shown to improve the performance of speaker dependent systems [10]. Figure 1 shows the group delay for anger and neutral for the same phoneme uttered by two speakers and the difference between the two emotions can be seen. However, it can also be seen that group delay varies significantly between the two speakers. The first 10 coefficients of the DCT of group delay is used as the feature vector.

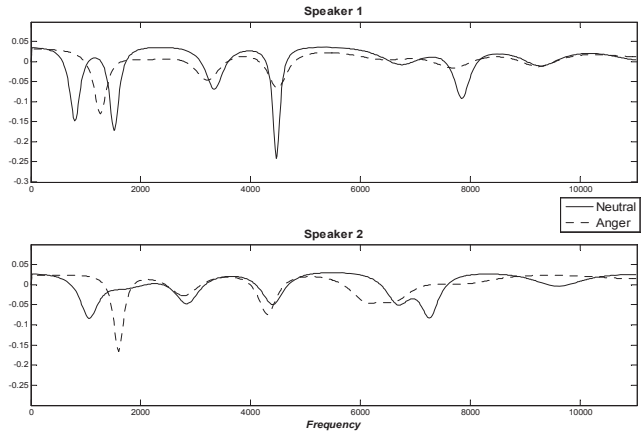

Fig. 1. Group delay for /a:/ for two emotions for two speakers.

2.3. Broad Spectral Measures (BSM)

It has been suggested that features characterising the vocal tract such as MFCCs and LPC based group delay are outperformed by vocal chord parameters such as pitch and energy in speaker independent emotion classification systems [4]. This is most likely due to the non-trivial differences in the vocal tract characterisations for different speakers. Thus, a feature vector that is derived from the speech spectrum, but excludes details that vary between different speakers will be useful for a speaker-independent emotion classifier. Features that characterise some aspect of the speech spectrum but do not describe it completely or in any detail are termed broad spectral measures in this work, and are typically low dimensional.

Energy Slope and Zero Crossing Rate (SZ) – Energy slope (sometimes referred to as spectral balance) is calculated as the ratio of the energy in the low frequency band (0-1 kHz) to that in the high frequency band (2-11 kHz). Zero crossing rate serves as a rough estimate of the dominant frequency present in the speech signal. Taken together they form a rough estimate of the spectral distribution of energy in the signal. In [4], they were proposed as additions to pitch and energy in a speaker independent system and are a 2 dimensional feature vector.

EMD Based Weighted Frequency – The recently pioneered empirical mode decomposition (EMD) can be used to represent the speech signal as a sum of zero-mean AM-FM components, which then allow for the definition of a positive instantaneous frequency for each component,

based on the Hilbert transform. A Weighted frequency feature based on these instantaneous frequencies has been recently proposed as an alternative to the energy slope feature and is a 3 dimensional vector [12].

$$wf[n] = \frac{\sum_{m=1}^{M} a_m[n]\theta_m[n]}{\sum_{m=1}^{M} a_m[n]} \quad (1)$$

Where $a_m[n]$ and $\theta_m[n]$ are the instantaneous amplitude and frequency of the m^{th} AM-FM component.

Spectral Centroid – One way to condense the information contained in the speech spectrum is to obtain a broad measure of the spectral magnitude distribution, such as spectral centroid. In this work, spectral centroid in each frame is single dimensional and was computed as follows:

$$spectral\_centroid = \frac{\sum_{i=1}^{N} |X(i)| \cdot F_s \cdot i}{N \sum_{i=1}^{N} |X(i)|} \quad (2)$$

Where, N is the frame size, $X(k)$ is the DFT of the framed signal and F_s is the sampling rate.

3. CLASSIFICATION SYSTEM

3.1. Speaker Normalisation

Previously, we used a modified feature warping technique as a means of speaker normalisation [9]. We apply it also to features in selected experiments reported in this paper.

3.2. Back-End

Sequential classifiers such as HMM-based classifiers have been advocated as being better suited for the task of emotion classification than other classifiers [4]. As an alternative, the feature vector can be modified to include temporal information and used with a non-sequential classifier such as probabilistic neural networks [10]. While the first approach was found to be suitable for a speaker-independent system, the smaller data set available for training in the case of a speaker-dependent system means that probabilistic neural networks are able to generalise better in the latter case.

For the purposes of this work however, where speaker-independent and speaker-dependent systems are to be compared with each other, a consistent classification setup is necessary. The size of the training dataset for the speaker dependent system was too small to train HMMs while probabilistic neural networks required extremely large amounts of system resources when used in a speaker independent system. Thus, a GMM-based classifier that could be trained on both datasets and makes a decision on a frame-by-frame basis was chosen for the experiments. Preliminary informal experiments indicate that the performance of the GMM-based classifier is almost as good as the optimal classifiers. In order to classify a test utterance, the log likelihoods of all frames belonging to that utterance were added and a maximum likelihood decision was made based on the summed values.

4. EXPERIMENTS

For our investigation, we used the LDC Emotional Prosody Speech corpus, comprising speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers. There is therefore no semantic or contextual information available. The entire database consists of 7 actors expressing 15 emotions for around 10 utterances each. When recording the database, actors were instructed to repeat a phrase as many times as necessary until they were satisfied the emotion was expressed and then move onto the next phrase. Only the final repetition of each phrase was used in this experiment.

Experiments for a five-emotion classification problem involving Neutral, Anger, Happiness, Sadness and Boredom were performed using a GMM based classifier, implemented in both speaker dependent and speaker independent configurations, using all features described in Section 2. All speaker dependent experiments were repeated 7 times, using 60% of the phrases from each of the 7 speakers as the training data set and the other 40% as the test data set, while the speaker independent experiments were repeated 7 times in a 'leave-one-out' manner, using data from each of the 7 speakers as the test set in turn and the data from the other 6 as the training set. In both cases, the accuracies reported are the means of the seven trials in each experiment.

Table 1. Comparison of five-class emotion classifier accuracies using various individual features

Features	Classification Accuracy		
	Spk Dep.	Spk Indep.	
		No Warp	Warp
Pitch + Energy (PE)	56.1 %	38.9 %	46.6 %
Energy Slope + ZCR (SZ)	55.4 %	34.9 %	46.5 %
Weighted Frequency (WF)	59.0 %	38.9 %	**48.9 %**
Spectral Centroid (SC)	51.8 %	34.4 %	39.2 %
MFCC	**74.1 %**	42.6 %	37.6 %
Group Delay (GD)	67.6 %	37.3 %	37.8 %
Formants (F)	51.1 %	34.9 %	42.6 %

As can be seen from these accuracies, the best performing features for the speaker dependent and independent systems differ. While the pitch and energy perform reasonably consistently in both configurations, the performances of spectral features are more interesting. The detailed spectral measures, such as MFCCs and the LPC based group delay perform well in speaker dependent systems. However, they are not very useful in the speaker independent system even though the information contained in them is similar to that in the formants. This tends to

suggest that detailed spectral measures such as MFCCs and group delay, while being able to distinguish between emotions well, also characterise the speaker to a much larger extent than both the speech production cues and the broad spectral measures.

Moreover, it can be seen that feature warping based speaker normalisation does not appear to improve the performance of group delay and in fact MFCCs perform better without warping. The broad spectral measures on the other hand do not perform very well when used alone. Hence, systems using combinations of excitation cues, broad spectral measures (BSM) and detailed spectral measures were compared and the accuracies are reported in Table 2.

The grouping of MFCCs with broad spectral measures provides the most effective feature combinations for the speaker dependent system. However, when compared with the MFCC alone speaker dependent system, the improvement due to the broad spectral measures are very marginal. This suggests that rather than them providing additional information, the speech production cues seem not as effective as MFCCs, and can even reduce the accuracy of the emotion models when combined with MFCCs.

Table 2. Five-class emotion classifier accuracies for various feature-pair combinations

Features	Classification Accuracy		
	Spk Dep.	Spk Indep.	
		No Warp	Warp
PE + SZ	56.8 %	39.4 %	53.7 %
PE + WF	64.7 %	39.2 %	53.2 %
PE + SC	60.4 %	40.2 %	45.7 %
PE + MFCC	69.8 %	40.5 %	50.0 %
PE + GD	69.8 %	45.2 %	52.4 %
PE + F	52.5 %	39.1 %	**59.5 %**
MFCC + SZ	74.1 %	41.5 %	44.7 %
MFCC + WF	**74.8 %**	45.8 %	42.6 %
MFCC + SC	**74.8 %**	42.1 %	41.5 %
PE + SZ + MFCC	69.8 %	45.5 %	50.8 %
PE + SZ + GD	70.5 %	48.5 %	52.9 %
PE + SZ + F	61.2 %	39.7 %	56.9 %

Among the speaker independent systems, the best performing features are the speech production cues combining the pitch and energy with the first three formants. The accuracy of the system for the combined front-end is significantly greater than the accuracies of the systems with the pitch and energy or formant information on their own. This strongly suggests that these two features are complementary, which is not surprising considering that formants are determined by vocal tract resonances while the glottal excitation cues characterise the vocal chords. Front-ends combining MFCCs with other features in Table 2 do not perform very well in the speaker independent case. The low overall recognition accuracies of all the front ends also indicate that these systems are not sufficient for any stand alone practical emotion recognition system.

5. CONCLUSION

This paper has compared features and feature pairs from three broad groups of feature types, for the purpose of emotion classification. The accuracies of the different features in a five-class emotion classification reported in this paper suggests that MFCCs are very discriminative but are also very characteristic of the speaker, and that they do not lend themselves well to speaker normalisation. Since most practical emotion classification systems would need to be speaker independent, MFCCs may not be the front-end of choice, unlike in speech recognition and speaker recognition systems. The comparisons also suggest that the optimum front-end for a speaker independent emotion recognition system is one that characterises both the vocal tract and the vocal chords consisting of the first three formant frequencies and energies along with excitation cues.

6. REFERENCES

[1] Salovey, P., Kokkonen, M., Lopes, P., and Mayer, J., "Emotional Intelligence: What do we know?", Manstead, A.S.R., Frijda, N.H., Fischer, A.H. (Eds.), *Feelings and Emotions: The Amsterdam Symposium*. Cambridge University Press, Cambridge, UK, pp. 321-340, 2004.

[2] Bhatti, M.W., Wang, Y., and Guan, L., "A Neural Network approach for Human Emotion Recognition in Speech", in *Proc. IEEE ISCAS*, pp. II- 181-184, 2004.

[3] Schuller, B., Rigoll, G., and Lang, M., "Hidden Markov Model based Speech emotion recognition", in *Proc. IEEE ICASSP*, vol. 2, pp. II- 1-4, 2003.

[4] Huang, R., and Ma, C., "Towards a Speaker-Independent Real-time Affect Detection System", in *Proc. 18th Int. Conf. on Pattern Recognition* (ICPR'06), vol. 1, pp. I- 1204-1207, 2006.

[5] S. Yacoub, S. Simske, X. Lin, J. Burns, "Recognition of emotions in interactive voice response systems," in *Proc. EUROSPEECH,* pp. 729-732, September 2003.

[6] Verceridis, D., Kotropoulus, C., and Pitas, I., "Automatic Emotional Speech Classification", in *Proc. IEEE ICASSP*, vol. 1, pp. I- 593-596, 2004.

[7] Wang, Y., and Guan, L., "An investigation of speech-based human emotion recognition", in *Proc. IEEE 6th Workshop on Multimedia Signal Processing,* pp. 15-18, 2004.

[8] Lugger, M., and Yang, B., "An incremental analysis of different feature groups in speaker independent emotion recognition", in *Proc. 16th Int. Congress of Phonetic Sciences,* pp. 2149-2152, 2007.

[9] Sethu, V., Ambikairajah, E., and Epps, J., "Speaker normalisation for speech based emotion detection," in *Proc. 15th Int. Conf. Digital Signal Processing,* pp. 611-614, 2007.

[10] Sethu, V., Ambikairajah, E., and Epps, J., "Group Delay Features for Emotion Detection," in *Proc. INTERSPEECH,* pp. 2273-2276, 2007.

[11] Sethu, V., Ambikairajah, E., and Epps, J., "Empirical Mode Decomposition based Weighted Frequency feature for speech-based emotion classification", *Proc. IEEE ICASSP,* 2008.