Group Delay Features for Emotion Detection

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

[1] School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney NSW 2052, Australia

[2] National Information Communication Technology (NICTA), Australian Technology Park, Eveleigh 1430, Australia

vidhyasaharan@gmail.com, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

10.21437/Interspeech.2007-617

Abstract

This paper focuses on speech based emotion classification utilizing acoustic data. The most commonly used acoustic features are pitch and energy, along with prosodic information like the rate of speech. We propose the use of a novel feature based on the phase response of an all-pole model of the vocal tract obtained from linear predictive coefficients (LPC), in addition to the aforementioned features. We compare this feature to other commonly used acoustic features based on classification accuracy. The back-end of our system employs a probabilistic neural network based classifier. Evaluations conducted on the LDC Emotional Prosody speech corpus indicate the proposed features are well suited to the task of emotion classification. The proposed features are able to provide a relative increase in classification accuracy of about 14% over established features when combined with them to form a larger feature vector.

Index Terms: emotion detection, group delay, linear predictive coefficients, probabilistic neural networks

1. Introduction

One significant difference between human-human interaction and human-machine interaction is the lack of emotional intelligence on part of the machine. Salovey *et al*. [1] defined emotional intelligence as having four branches: perception of emotion, emotions facilitating thought, understanding emotions and managing emotions. The focus of this paper is on the first of these four branches, to build a system that is able to detect the emotional state of a person based on speech. Automatic emotion detection of humans has a broad range of applications and as such has attracted considerable attention from researchers in the last few years. The range of applications extends from tele-monitoring of patients to computer based tutoring systems to call centre services where angry customers are automatically transferred to trained human representatives [2]. Depending on the application, the emotion classification system may have different requirements. For example, in the case of a system for tele-monitoring patients the target speaker is always the same person and the system can be speaker-dependent. However, a system designed for the aforementioned call centre application would have to be speaker-independent since every caller is a target speaker.

This paper looks at an emotion detection 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, hidden Markov models (HMM), Gaussian mixture models (GMM) and support vector machines (SVM) are used to detect the emotional state of the speaker [2-6]. Among the acoustic features used, those derived from pitch and energy, are the most popular. However both these features characterise only the vocal chords' state. A feature vector that can characterise the vocal tract as well as the vocal chords could further improve classification accuracy.

In this paper, we propose the use of a novel feature vector based on the all-pole filter model of speech production that forms the basis of LPC analysis. More specifically, the feature vector is extracted from the phase characteristics of the all-pole filter obtained from a speech segment.

Group delay has been used as a feature for phoneme recognition by Murthy *et al.* [7], however they compute it from the Fourier transform of the signal and not from the LPC coefficients. The advantage of the method proposed in this paper when compared to their approach is that the excitation component is separated from the vocal tract characteristics by the LPC algorithm.

2. Proposed Group Delay Feature

The all-pole filter model of speech production describes speech signal as the response of an all-pole when excited by either a pulse train (voiced speech) or random noise (unvoiced speech). The excitation signal characterises the vocal chords' while the filter characterises the vocal tract of the speaker. We believe the commonly used pitch and energy based features depend on the excitation signal and consequently characterise the state of the vocal chords. The features proposed in this paper are an effort to describe the state of the vocal tract. The magnitude responses of the all-pole filters provide important information about formant locations which determine the phoneme and have been used as features in speech recognition and speaker verification problems. However, the phase response has rarely been studied as a feature. We believe the group delay of the all-pole filter characterising the vocal tract contains information that can help determine the emotional state of the person.

The formant positions, which can be obtained from the magnitude response of the all-pole filter, determine the sound (phoneme) produced. However, when the same phoneme is uttered by a person in different emotional states, formant positions may not be very different. Figure 1 shows the formant positions obtained from the phoneme */a:/* in the word *'thousand'* which was uttered by the same person in different emotional states. It can be observed that while locations of the first three formants are not very different, the formant bandwidths are very different and produce the difference in the sounds that help distinguish between the two emotions. This change in bandwidth is reflected in the group delays of

the corresponding all-pole filters (shown in Figure 2). The magnitude of the group delay increases with a reduction in the formant bandwidth and the positions of the group delay's local minima reflect the formant locations.

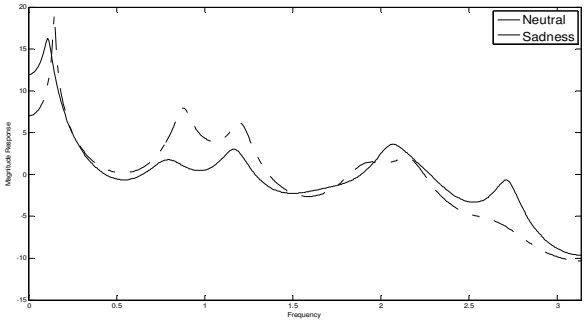

Figure 1: *Formant locations for /a:/ for two emotions spoken by the same person.*

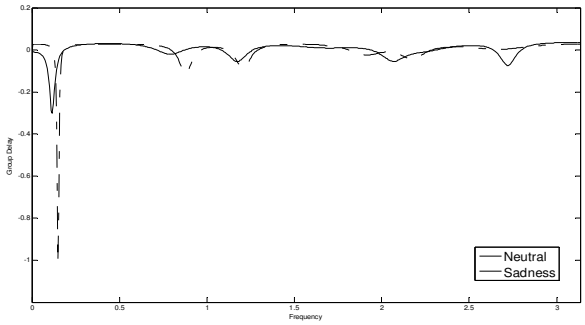

Figure 2: *Group delay for /a:/ for two emotions spoken by the same person.*

In order to see the relationship between formant bandwidths and the group delay value at the formant frequency, we examine the transfer function of the all-pole filter that characterises the vocal tract. It can be considered to be a cascade of second order resonators with conjugate poles, with each resonator producing a formant. Setting $z = e^{j\omega}$ gives the frequency response of the all-pole filter.

$$V(\omega) = \prod_{i=1}^{M} H_i(\omega) \tag{1}$$

$$|V(\omega)|e^{j\theta(\omega)} = \prod_{i=1}^{M} |H_i(\omega)|e^{j\phi_i(\omega)} \tag{2}$$

Where $V(\omega)$ is the transfer function of the all-pole vocal tract filter, $H_i(\omega)$ is the transfer function of the i^{th} 2-pole resonator which produces the i^{th} formant, $\omega \in [-\pi,\pi]$ and M is the total number of formants.

The group delay of the all-pole filter can thus be written as the sum of group delays of the resonators

$$\theta(\omega) = \sum_{i=1}^{M} \phi_i(\omega) \tag{3}$$

Thus, studying the relationship between the group delay value at the resonant frequency and the formant bandwidth for the 2-pole resonators should be adequate. In order to do so, consider the frequency response of a 2-pole resonator.

$$H_i(\omega) = \frac{1}{\left(1 - re^{j\alpha_i}e^{-j\omega}\right)\left(1 - re^{-j\alpha_i}e^{-j\omega}\right)} \tag{4}$$

Where α_i is the formant (resonant) frequency and $re^{\pm j\alpha_i}$ are the poles of the system. From (4), the squared magnitude response, and consequently the formant bandwidth are computed as follows:

$$|H_i(\omega)|^2 = \frac{1}{\left[1 + r^2 - 2r\cos(w - \alpha_i)\right]\cdot\left[1 + r^2 - 2r\cos(w + \alpha_i)\right]} \tag{5}$$

When the poles are near the unit circles, i.e., r is close to but less than 1, the formant bandwidth, $\Delta\omega$, can be approximated as

$$\Delta\omega \approx 2(1 - r) \tag{6}$$

Also from (4), the phase response of the system can be computed as:

$$\phi_i(\omega) = -\left[\tan^{-1}\left(\frac{r\sin(\omega - \alpha_i)}{1 - \cos(\omega - \alpha_i)}\right) + \tan^{-1}\left(\frac{r\sin(\omega + \alpha_i)}{1 - \cos(\omega + \alpha_i)}\right)\right] \tag{7}$$

The group delay is obtained by differentiating the phase response with respect to frequency. For the 2-pole resonator, the group delay obtained is as follows:

$$\tau_g(\omega) = \left[\begin{array}{l} \dfrac{r^2 - r\cos(\omega - \alpha_i)}{1 + r^2 - 2r\cos(\omega - \alpha_i)} \\ + \dfrac{r^2 - r\cos(\omega + \alpha_i)}{1 + r^2 - 2r\cos(\omega + \alpha_i)} \end{array}\right] \tag{8}$$

At the resonant frequency α_i, the group delay takes the following value:

$$\tau_g(\alpha_i) = \frac{-r}{1 - r} + \frac{r[1 - \cos(2\alpha_i)]}{1 + r^2 - 2r\cos(2\alpha_i)} \tag{9}$$

It can be seen that as the value of r approaches 1, the group delay function's value at the formant frequency takes an increasingly negative value since the magnitude of the first term in (9) is always larger than the magnitude of the second term for all $r > 0.2361$.

From equations (6) and (9) it can be seen that a reduction in the formant bandwidth is reflected by an increasingly larger negative value of the group delay at the formant frequency. Since the overall group delay of the all-pole filter is the sum of the group delays of the resonators, we can expect the group delay to have negative spikes at formant locations, with the magnitudes of these spikes reflecting the formant bandwidths.

In order to estimate the group delay, the all-pole filter parameters are estimated using the LPC algorithm. For our experiments we used a 15 pole filter to model the vocal tract. The phase response of this filter is estimated from the first 1024 samples of the impulse response and the group delay is calculated by differentiating this phase response with respect to frequency. Alternately, equation (9) gives the contribution of each complex conjugate pole pair to the overall group delay, which can thus be estimated by adding the contributions of all the poles present in the vocal tract filter.

The group delay computed this way is a vector with a large number of components. In order to represent it compactly, we perform a discrete cosine transform (DCT) on the sequence and pick the first 10 coefficients ($\beta_0 - \beta_9$) as the elements of our feature vector. The following figure shows the steps involved in computing the LPC group delay feature vector.

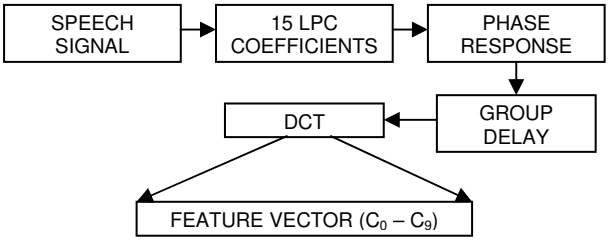

Figure 3: *LPC group delay feature extraction.*

3. The Emotion Detection System

3.1. The Front-End

For our system (Fig. 2) we use the LPC group delay feature vector along with the features used by Huang *et al.* [6]. Thus, the 10 group delay features were concatenated with pitch, energy, zero crossing rate (ZCR) and energy slope to give a total of 14 features per frame. The YIN estimator [8] was used to estimate pitch. Similar to the definition in [6], the energy slope was calculated as the ratio of the energy contained in the low frequency band (0–1 kHz) to that in the higher frequency band (2 – 11 kHz; a sampling rate of 22 kHz was used in this study).

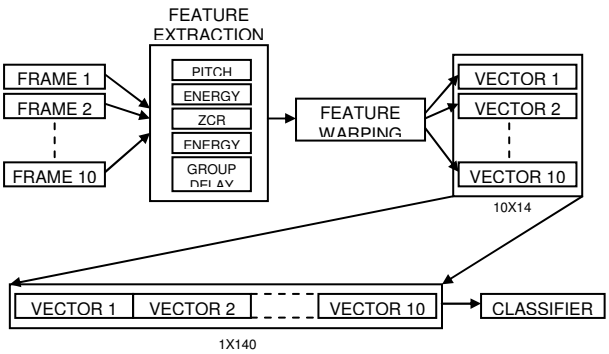

Figure 4: *System overview*

All features were computed within frames of 40ms duration (minimum duration for reliable pitch estimation) obtained using a rectangular window with consecutive frames overlapping by 30 ms. The 14 dimensional feature vectors estimated from 10 consecutive frames were then concatenated to form a larger feature vector which was then passed to the back-end. This concatenation of features from 10 consecutive frames was done so that information contained in temporal variations of the features was taken into account by the non-sequential classifier used in the back-end of our system. Thus the classifier makes a decision based on 130ms of speech.

3.2. Feature Warping

Feature warping, or cumulative distribution mapping, is a technique that maps each feature to a predetermined distribution, originally suggested as a method to provide robustness against channel mismatch and non-linear noise effects [9]. It has also been suggested that warping the features to a normal distribution provides better matching to GMM-based back-ends [10]. In our preliminary studies we observed that feature warping resulted in a slightly improved clustering of data in a feature space and was able to improve the performance of any classifier. For our system, feature warping improves the accuracy by about 1%-4% (depending on the feature used).

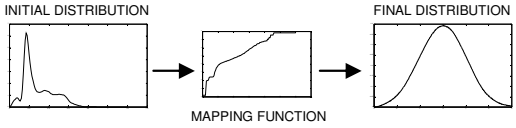

Figure 5: *Overview of feature warping*

3.3. The Back-End

Any one of the numerous available classifiers can be used at the back end of an emotion detections system. However it has been suggested that sequential classifiers (such as HMM-based classifiers) are better suited to this task as they take in account temporal variations in the features [6]. An alternative may be to modify the feature vector to contain temporal information and use a non-sequential classifier. In this paper we follow the second approach and use a probabilistic neural network (PNN) as our back end. An important reason for this choice is that in our preliminary studies we found that a PNN is able to generalise better when using a smaller data set for training as opposed to HMM or GMM based classifiers. While it is generally true a GMM based classifier can train on a much smaller database when compared to a HMM based one, in our case the GMMs would model a 140 dimensional feature space as opposed to the HMMs, which would model only a 14 dimensional feature space (Figure 4). Consequently the GMMs would require a large number of training vectors to reliably estimate the feature distributions. This assumption was borne out by our preliminary studies where using a PNN resulted in a higher accuracy when compared to both GMM and HMM based classifiers.

4. Experiments

For our experiments we used the LDC Emotional Prosody Speech corpus [11]. It consists of 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. When recording the database, the 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 last instance of each phrase was selected for this experiment.

The system described in section 3 (Figure 4) was implemented with different features in order to judge the performance of the proposed features. All 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 for a speaker dependent system. Experiments for a five-emotion classification problem involving Neutral, Anger, Happiness, Sadness and Boredom were performed.

Four different feature sets were used in all experiments in order to do a comparative study. The first one consisted of the same features used by Huang *et al.* [6] namely, pitch, energy,

zero-crossing rate and energy slope (ZEPS). The second feature set was the 10 dimensional LPC group delay feature proposed in this paper. The third feature set was a combination of the first two obtained by concatenating them to form a 14 dimensional vector, and the final feature set used was a vector composed of 12 Mel frequency cepstral coefficients. For all the experiments, the phrases were divided into sequences of 10 consecutive frames (each of duration 40ms with a 30ms overlap) and each sequence was evaluated independently in order to facilitate scoring. Feature warping was used in all cases. The results of these experiments are given in Tables 1 and 2 below.

Table 1. *Emotion classification accuracy for the five-class problem.*

Test Speaker	ZEPS	Group Delay (GD)	ZEPS + GD	MFCC (12)
1	**35.5%**	28.4%	32.4%	29.2%
2	49.8%	53.0%	**58.9%**	50.6%
3	**47.0%**	35.1%	43.3%	45.7%
4	53.5%	56.7%	**65.7%**	54.5%
5	70.5%	68.7%	**81.6%**	76.0%
6	63.2%	72.3%	**78.9%**	70.9%
7	40.1%	40.8%	**50.3%**	45.2%
Mean	*51.4%*	*50.7%*	***58.7%***	*53.2%*

It can be seen that the LPC group delay features, when added to the ZEPS feature set proposed by Huang *et al.* [6], give the highest accuracy for the five-emotion classification problem. We found that combining the MFCCs with the ZEPS feature set does not provide as significant an improvement as the group delay (55% as opposed to 59%). This is probably because MFCCs and ZEPS are more correlated than ZEPS and group delay. The following table lists the average accuracies (across all 7 speakers) for the five emotions.

Table 2. *Average class accuracy for five-emotion classifier.*

Emotion	ZEPS	Group Delay (GD)	ZEPS + GD	MFCC (12)
Neutral	69.0%	74.2%	**84.8%**	74.7%
Anger	52.0%	43.9%	**57.6%**	53.5%
Sadness	31.2%	41.7%	**48.6%**	52.2%
Happiness	49.6%	43.5%	**52.3%**	45.4%
Boredom	44.2%	42.7%	**49.7%**	42.8%
Mean	*51.4%*	*50.7%*	***58.7%***	*53.2%*

From this table it is clear that the combination of ZEPS feature set [6] and the proposed LPC group delay features exhibit the highest accuracies for each of the five emotions. Moreover none of the accuracies are significantly lower than the other four and all of them are much higher than the accuracy for random guessing (20%). This indicates the proposed features are able to characterise all five emotions reasonably well.

5. Conclusion

This paper presents a novel feature to increase the accuracy a multi-class emotion detection system. We estimate the group delay of the all-pole filter from its phase response. A discrete cosine transform is then used to represent this group delay compactly as a feature vector. The group delay is able to characterise both formant locations and formant bandwidths and thus provides a reasonable good model of the vocal tract state of the speaker which in turn is dependent on his or her emotional state. The results included in the paper show that for a five emotion classification problem, the addition of the proposed features results in a relative increase in accuracy of about 14% over established features. Preliminary tests have indicated that the proposed features are not as effective for a speaker independent system where training and testing data are from different speakers as they are for a speaker dependent system. We believe this is because the difference between vocal tract parameters of different speakers is much greater and hence overshadows the variations caused due to change in emotional state. Currently work is underway on an attempt to normalise this feature between speakers so as to use it in a speaker independent system.

6. Acknowledgements

This research was fully funded by National Information and Communication Technology, Australia (NICTA).

7. 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] Yacoub, S., Simske, S., Lin, X., and Burns, J., "Recognition of Emotions in Interactive Voice Response systems", in *Proc. EUROSPEECH*, pp. 729-732, 2003

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

[4] 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

[5] 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

[6] 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

[7] Murthy, H.A. and Gadde, V., "The Modified Group Delay Function and its application to Phoneme Recognition", in *Proc. IEEE ICASSP*, pp. I- 68-71, 2003

[8] A. de Cheveigne, and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", *JASA*, vol. 111, Issue 4, pp. 1917-1930, 2002

[9] Pelecanos, J., and Sridharan, S., "Feature warping for robust speaker verification", in *Proc. A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 243-248, 2001

[10] Allen, F., Ambikairajah, E., and Epps, J., "Language Identification using Warping and the Shifted Delta Cepstrum", in *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, 2005

[11] Emotional Prosody Speech corpus, Linguistic Data Consortium, University of Pennsylvania, PA, USA, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28