SPEAKER NORMALISATION FOR SPEECH-BASED EMOTION DETECTION

Vidhyasaharan Sethu[1,2], *Eliathamby Ambikairajah*[1,2] *and Julien Epps*[3,1]
vidhyasaharan@gmail.com, ambi@ee.unsw.edu.au, j.epps@unswasia.edu.sg

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

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

[3]UNSW Asia, 1 Kay Siang Road, Singapore 248922

ABSTRACT

The focus of this paper is on speech-based emotion detection utilising only acoustic data, i.e. without using any linguistic or semantic information. However, this approach in general suffers from the fact that acoustic data is speaker-dependent, and can result in inefficient estimation of the statistics modelled by classifiers such as hidden Markov models (HMMs) and Gaussian mixture models (GMMs). We propose the use of speaker-specific feature warping as a means of normalising acoustic features to overcome the problem of speaker dependency. In this paper we compare the performance of a system that uses feature warping to one that does not. The back-end employs an HMM-based classifier that captures the temporal variations of the feature vectors by modelling them as transitions between different states. Evaluations conducted on the LDC Emotional Prosody speech corpus reveal a relative increase in classification accuracy of up to 20%.

Index Terms— Feature warping, cumulative distribution mapping, emotion detection, hidden Markov model

1. INTRODUCTION

Humans express emotions through various means, including speech, facial expressions, gestures, and other non-verbal cues. Humans are also, in most cases, able to recognise the emotion being conveyed and react accordingly so as to ensure the interaction amongst them is successful. Over the past few years, researchers have turned their attention to improving human-machine interaction by making it similar to human-human interaction, with the addition of automatic emotion detection. Automatic detection of human emotional states is a technology with a broad range of applications, and as such has attracted considerable attention from researchers in the last five years. The range of applications for automatic emotion detection extends from call centre services, where a machine transfers calls from angry customers to human representatives [1], to computer-based tutoring systems, to medical telemonitoring of patients.

Pitch, intensity, speech rate and various spectral features constitute the basis of most recent techniques proposed for emotion detection that utilise acoustic signals. Based on these measures, the emotion detection system then uses classifiers such as neural networks, support vector machines, Gaussian mixture models, hidden Markov models and decision trees to recognise the emotion being conveyed by the speaker [1-5]. Among the acoustic features used for emotion classification, pitch and energy are the most popular. These features, however, are also dependent on the speaker and can exhibit a lot of variability between speakers. This would not be a problem for a classification system that is trained on data obtained from the target speaker, but such an expectation is not practical in most cases.

In this paper, we examine the problem of recognising and classifying emotions based on conversational human speech composed of short utterances. In particular, the use of feature warping as a means of reducing the abovementioned variation between speakers is examined, with a view to improving the performance of the classifier.

2. THE EMOTION RECOGNITION SYSTEM

2.1. Classification

Sequential classifiers (such as HMM-based classifiers) have been suggested to be better suited for the task of classifying emotions than other commonly used non-sequential classifiers such as support vector machines and decision trees [5]. Observations from our preliminary investigations into back-end configurations tend to agree with this suggestion. An alternative may be the inclusion of temporal information into the feature vector in the form of delta or shifted delta features [4]. In this paper we use an HMM-based classifier and a short feature vector devoid of delta features similar to the system used in [5]. The use of a short feature vector makes it easier to study the effect of speaker normalisation on each feature, while the use of an HMM-

based classifier retains temporal information. Each state of the HMM is modelled by a Gaussian mixture model (GMM). For each emotion, a hidden Markov model is trained and the emotion corresponding to the model that best matches the incoming test feature sequence is chosen as the detected emotion. In contrast to the system used in [5], we use shorter feature sequences in order to increase the number of tests, enabling us to estimate reliable statistics at the cost of slightly reduced accuracy. We do so since our aim is to better illustrate the effect of speaker normalisation. In our experiments all emotions were modelled by 4 state HMMs with each state represented by a GMM containing 4 mixtures. Empirical work indicated this configuration provided the best trade off between generalisation and accurate modelling of the feature distributions.

2.2. Feature Extraction

For our system (Fig. 1) we chose the same feature set as Huang *et al*. [5], and used a four-dimensional feature vector composed of pitch, energy, zero crossing rate (ZCR) and energy slope. The YIN estimator [6] was used to estimate pitch. Similar to the definition in [5], the energy slope was calculated as the ratio between 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). All 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. All groups of 10 consecutive frames were then taken as training and testing sequences for the HMMs. Empirical tests indicated that 10 frames was the minimum sequence length required for good accuracy. Thus, each emotion model was trained using sequences of 10 feature vectors computed from 130ms of emotional speech. Pitch estimates were not available for all frames (pitch is not estimated for unvoiced speech), therefore only sequences where a pitch estimate was available for all 10 frames were used for both training and testing phases.

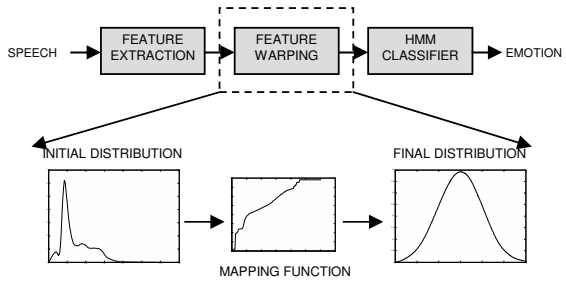

Figure 1: System Overview

3. FEATURE WARPING

Feature warping 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 [7–8]. Also known as histogram equalisation in image processing literature and cumulative distribution mapping in speech processing literature, feature warping has been used successfully in speech recognition [8], speaker verification [7] and language identification [9].

Feature warping treats each feature as an independent stream of values, mapping them onto a target distribution. Denoting the target distribution as $h(z)$, and the original probability distribution of the feature as $f(y)$, the mapping is defined as

$$\int_{y=-\infty}^{p} f(y)dy = \int_{z=-\infty}^{q} h(z)dz, \qquad (1)$$

where p is the original feature value and q is the warped feature value. It is not necessary; however, to estimate the actual distribution $f(y)$, rather the following procedure is used to map the feature values onto a target distribution.

For each value in a given feature stream of length N:

(1) The ranking R is calculated as the new index of the value when all the features in the stream are sorted into descending order and indexed from 1 to N

(2) R is then used to obtain a mapped value q from a pre-calculated look table in which the values are calculated:

$$\frac{N+\beta-R}{N} = \int_{z=-\infty}^{q} h(z)dz, \quad (0<\beta<1) \qquad (2)$$

Recognising the integral on the right as the cumulative distribution function (CDF) and denoting the inverse CDF as $H^{-1}(x)$:

$$q = H^{-1}\left(\frac{N+\beta-R}{N}\right) \qquad (3)$$

3.1. Speaker Normalisation

Pitch is a widely used as a successful feature in emotion classification problems. Pitch values, however, exhibit a large amount of variation between speakers. Fig. 2(a-b) shows the probability distributions of the pitch values for two different speakers expressing no emotion (neutral) and anger (tagged as 'hot anger' in the LDC Emotional Prosody database [10]), estimated from all utterances from these two speakers present in the database. It is clear that while the distributions for neutral and anger are distinct for the speakers, the distributions for speaker 1 are not the same as

the distributions for speaker 2, in particular the neutral emotion in this example. Hence, when the probability distribution of pitch for an emotion is estimated for all speakers (in our case by the GMMs representing each state of the HMM of that emotion), the resultant distribution is multi-modal with a large variance (Fig. 2c).

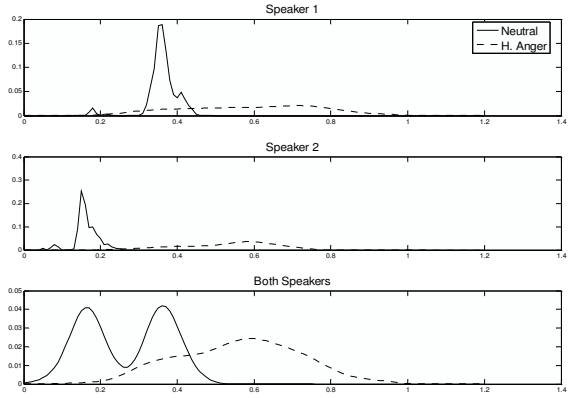

Figure 2: Distribution of pitch without feature warping
(a) Speaker 1 (b) Speaker 2 (c) Both Speakers

Feature warping maps a feature stream (pitch values in this case) onto any target distribution. Choosing the standard normal distribution as our target distribution, we perform feature warping on all data consisting of all emotions taken together for each speaker, independent of each other. That is, the overall distribution of the pitch stream (taking into account both emotions) for each speaker is mapped to the standard normal distribution. This preserves the difference between distributions for each emotion for a speaker while normalising the values across speakers. When used live, reliable feature warping of the test speaker can be accomplished by estimating the speaker statistics over time, as long as the same person remains in front of the microphone. An alternative when a large training set is not available would be to use data from a training speaker who is acoustically close to the test speaker, to achieve the desired warping.

The distributions estimated from the pitch streams for both speakers after feature warping are shown in Fig. 3(a-b). It can be seen that the variation between the distributions for both speakers is now much lower. This also results in a reduction in the variance of the overall distribution for an emotion estimated from both speakers (Fig. 3c).

4. EXPERIMENTS

For our experiments we used the LDC Emotional Prosody Speech corpus [10]. 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.

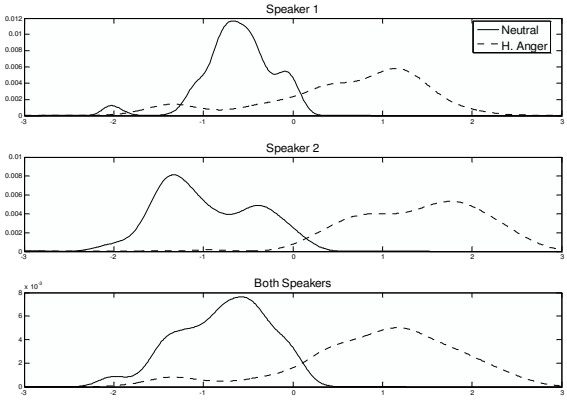

Figure 3: Distribution of pitch after feature warping
(a) Speaker 1 (b) Speaker 2 (c) Both Speakers

The system described in Section 2 (Fig. 1) was implemented both with and without feature warping for purposes of comparison. The 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. The results for the Neutral vs. Anger tests are shown below (Table 1). This test was chosen because anger detection is relevant for call centre applications and an important part of patient monitoring. For this experiment, the phrases were divided into sequences of 10 consecutive frames and each sequence was evaluated independently in order to facilitate scoring.

Table 1: Emotion classification accuracy for Neutral vs. Anger

Test Speaker	**Without Warping**			**With Warping**		
	Neutral	Anger	Overall	Neutral	Anger	Overall
1	82.8%	77.4%	79.1%	82.1%	**82.4%**	**82.3%**
2	99.1%	100%	99.4%	**100%**	100%	**100%**
3	96.2%	80.0%	89.6%	93.6%	**97.8%**	**95.3%**
4	52.1%	99.1%	79.0%	**100%**	85.1%	**91.5%**
5	100%	100%	100%	100%	100%	100%
6	100%	91.7%	94.1%	100%	**96.4%**	**97.5%**
7	100%	99.3%	99.7%	100%	**100%**	**100%**
Mean	*90.0%*	*92.5%*	*91.6%*	***96.5%***	***94.5%***	***95.2%***

These results show that when no feature warping is used, the accuracies for a few speakers are very high while those of

the remainder are low. This is due to the estimated statistics matching some speakers but not others. Speaker-specific feature warping directly addresses this deficiency (as described in Section 3), and results in improved matching between the estimated model and the data from all speakers.

A five emotion classification experiment was also performed to gauge the performance of speaker-specific feature warping for a more typical emotion classification problem. The five emotions selected for this test were neutral, anger (hot), happiness, sadness and boredom. These emotions were chosen as they are representative of the available emotions and are able to illustrate the limitations of feature warping. The results are shown below (Table 2).

Table 2: Five class emotion classification accuracy

Test Speaker	Without Warping	With Warping
1	26.7 %	**33.4 %**
2	34.7 %	**39.2 %**
3	34.5 %	**38.3 %**
4	28.8 %	**41.4 %**
5	40.2 %	**61.7 %**
6	52.4 %	40.3 %
7	29.6 %	**37.3 %**
Mean	*35.3 %*	***41.6 %***

For the 5-class problem, the classification accuracy is not consistent even after feature warping. In fact, it results in a significant loss of accuracy for speaker 6. This is because the features are not sufficiently discriminative for five classes. Fig. 4 shows the probability distribution of pitch values for all five emotions for one speaker before and after warping. It can be seen that there is significant overlap between the distributions, resulting in low accuracy. This effect outweighs any improvement that can be obtained by normalising between speakers by feature warping.

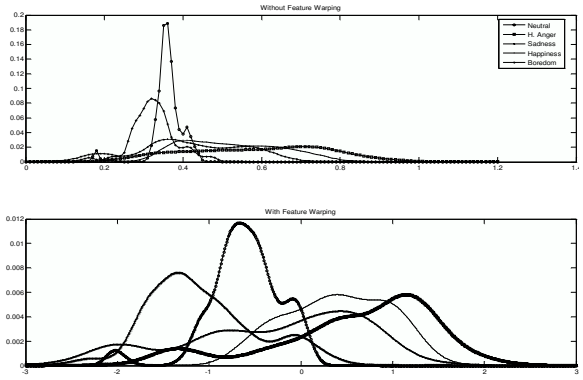

Figure 4: Distribution of pitch for the five class problem
a) Without warping b) With warping

5. CONCLUSION

This paper presents a technique to reduce the variance in data that arises due to differences in speaker characteristics, in order to improve the performance of a speaker independent emotion classification system. We apply feature warping to transform the data from each speaker such that they are all mapped to the same distribution. This results in the data being separated in the feature space according to emotional class but not individual speakers. The results included in this paper show that provided the features can intrinsically provide sufficient discrimination between classes for each speaker, feature warping can improve performance by normalising between speakers.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

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

[2] D. Verceridis, C. Kotropoulus and I. Pitas, "Automatic emotional speech classification," *Proc. IEEE ICASSP,* vol. 1, pp. I- 593-596, May 2004

[3] M. W. Bhatti, Y. Wang and L. Guan, "A neural network approach for human emotion recognition in speech," in *Proc. IEEE ISCAS,* vol. 2, pp. II- 181-184, May 2004

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

[5] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *Proc. 18th Int. Conf. on Pattern Recognition, (ICPR'06),* vol. 1, pp. 1204-1207, 2006

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

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

[8] A. Torre, J. Segura, C. Benitez, A. M. Peinado and A. J. Rubio, "Non-linear transformation of the feature space for robust speech recognition," in *Proc.IEEE ICASSP,* vol. 1, pp. I- 401-404, May 2002

[9] F. Allen, E. Ambikairajah and J. Epps, "Warped magnitude and phase based features for language identification," in *Proc. IEEE ICASSP,* vol. 1, pp. I- 201-204, May 2006

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