# Phonetic and Speaker Variations in Automatic Emotion Classification

*Vidhyasaharan Sethu* [1,2], *Eliathamby Ambikairajah* [1,2] *and 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

## Abstract

The speech signal contains information that characterises the speaker and the phonetic content, together with the emotion being expressed. This paper looks at the effect of this speaker- and phoneme-specific information on speech-based automatic emotion classification. The performances of a classification system using established acoustic and prosodic features for different phonemes are compared, in both speaker-dependent and speaker-independent modes, using the LDC Emotional Prosody speech corpus. Results from these evaluations indicate that speaker variability is more significant than phonetic variations. They also suggest that some phonemes are easier to classify than others.

**Index Terms**: emotion classification, speaker normalisation, phoneme recognition

## 1. Introduction

Human speech is probably the most natural and widely used form of interpersonal communication. Apart from the verbal communication that makes use of language and words to convey the bulk of the information in most cases, speech also contains other information, namely cues that identify the speaker and cues that express emotions. Researchers have been studying speech recognition and speaker identification for a number of years, and in recent years have also turned their attention to emotion recognition, due to a broad range of potential applications. All three are necessary to improve human-machine interactions and bring them closer to human-human interactions.

The focus of this paper is on how phonetic and speaker specific cues affect a speech based emotion classification system. The system considered here does not make use of semantic or linguistic information, and as such does not make use of language models. Such systems rely solely on prosodic and/or spectral features such as pitch, intensity, speech rate, cepstral coefficients, group delay, instantaneous frequency [1-6]. Based on these measures, the emotion classification system then uses classifiers such as neural networks, Gaussian mixture models, hidden Markov models, support vector machines and decision trees to recognise the emotion being conveyed by the speaker [1-8]. Among these the acoustic and prosodic features such as pitch, intensity and speech rate can exhibit considerable variability between different speakers, while spectral features such as cepstral coefficients and group delay can exhibit significant variability between both different phonemes and different speakers.

Previously, we showed that a modified feature warping technique can be used to reduce inter-speaker variability and improve the accuracy of a speaker independent emotional classifier [8].

In this paper, we examine the classification accuracies of such a speaker independent system (using data from different speakers for training and testing) and a speaker dependent system (using data from the same speaker for training and testing) for different phonemes with a view to improving the performance of the classifier. The aim of this paper is to determine whether some phonemes are more conducive to emotion classification than others, and then to examine the variability between different speakers, rather than to build the optimal classification system. In order to achieve this, an emotion classifier is setup and the independent classification accuracies for different phonemes are determined. If certain phonemes express the emotion being conveyed better than others, the classification accuracies of those phonemes should be correspondingly higher than those of other phonemes.

The classifier used in the experiments reported in this paper addresses a five-class problem, classifying each spoken utterance into one of the five emotions contained in the LDC Emotional Prosody corpus: anger, sadness, happiness, boredom and neutral (no emotion).

## 2. Emotion Recognition System

### 2.1. Feature Extraction

For the purpose of the comparisons we chose the feature set proposed by Huang *et al*. [4], namely, pitch, energy, zero crossing rate and energy slope. The YIN estimator [9] was used to estimate pitch. Energy slope was 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; a sampling rate of 22 kHz was used in this study). All features were computed within frames of 40ms duration obtained using a rectangular window with consecutive frames overlapping by 30ms. Pitch estimates were not available for all frames, and only those with reliable pitch estimates were used, for training and testing. Previously, we have proposed the use of group delay based features for speaker dependent systems [5] and instantaneous frequency based features [6] for speaker independent systems. These features are however not used in this study since the aim is not to build the optimum classifier and dropping these features simplifies the system and makes the results easier to interpret.

### 2.2. Feature Warping

Feature warping, or cumulative distribution mapping, is a technique that maps each feature to a predetermined distribution (Fig. 1), originally suggested as a method to provide robustness against channel mismatch and non-linear noise effects [10]. Previously, we have used a modified feature warping technique as a means of speaker normalisation [8] and applied it to the features in some of the experiments reported in this paper. Even though speaker normalisation is redundant in a system that is trained and tested on data from the same speaker, feature warping can still be applied in order to be consistent with the other experiments to which it is compared. It has also been suggested that warping the features to a normal distribution provides better matching to GMM-based back-ends [11].

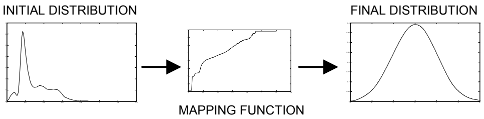

Figure 1: *Overview of Feature Warping*

### 2.3. Classification

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 [5]. 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 to each other, a consistent classification setup is necessary. Thus, a GMM-based classifier that makes a decision on a frame by frame basis was chosen for the experiments. Preliminary informal experiments indicated that while this resulted in a reduction in the overall accuracy of the system, the trends observed over the different parameters being analysed were consistent with those observed when the optimal classifiers were used.

### 2.4. Phoneme Recognition

In order to examine the effect of phonetic content on classifier performance it is essential to determine the phoneme associated with every frame of data. The phoneme recogniser developed at the Faculty of Information Technology, Brno University of Technology [12-13] was applied to generate phonetic labels from the data. Informal tests on the TIMIT database indicate the phone recogniser has an accuracy of about 74%. The dominant phoneme in each frame (the phoneme with the longest duration in the frame when more than one was present) according to the labels was then associated with the frame, as seen in Fig. 2.

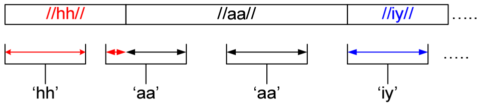

Figure 2: *Frame-level phonetic labelling*

The phoneme set consists of 39 phonemes, as described in [12]. However, since reliable pitch estimation is rarely possible from stops, affricates and fricatives, they were all combined as a single phoneme group (the phonemes *b, d, g, p, t, k, dx, jh, ch, s, sh, z, f, th, v,* and *dh* were grouped together and labelled as *fr*). Also, frames labelled as silences or pauses were not included in the experiments. This gave a total of 23 classes.

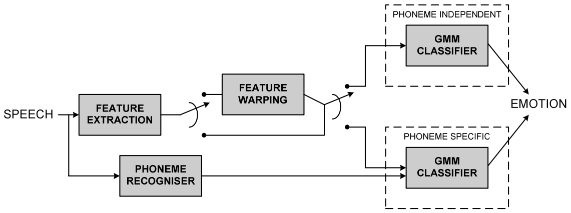

Figure 3: *System Overview*

## 3. Experiments

For our experiments we used the LDC Emotional Prosody Speech Corpus [13], comprising of speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers in order to ensure no semantic or contextual information is available. The entire database consists of 7 actors expressing 15 emotions for around 10 utterances each. 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 instances of the phrases were used in these experiments.

The system described in section 2 (Fig. 3) was implemented and the classification accuracies for all the phoneme classes listed in Table 1 were determined for speaker dependent and speaker independent cases. For each emotion, a Gaussian mixture model was trained and maximum likelihood estimation was used to determine the emotional class of the test data. All the GMMs used in this experiment contain 16 mixtures. Empirical work indicated this configuration provided a good trade-off between generalisation and accurate modelling of the feature distribution. For the speaker-dependent experiments, 70% of the data from a speaker was used as training data and the remaining 30% as the test data. This was repeated for all 7 speakers and the average accuracies per phoneme class were calculated. The speaker-independent experiments were repeated 7 times in a 'leave-one-out' manner using data from each of the 7 speakers as test data and that from the other 6 as training data. Once again, the average phonetic class accuracies were computed from all 7 trials. The number of frames of test data available for each phonetic class for both speaker dependent and speaker independent experiments are listed in Table 1.

Table 1. *Number of Test Frames in each Phonetic Class*

| Phonemes | No. of Frames (Speaker Dependent) | No. of Frames (Speaker Independent) |
|---|---|---|
| *fr* | 526 | 1911 |
| *m* | 232 | 906 |
| *n* | 624 | 2413 |
| *ng* | 10 | 40 |
| *l* | 583 | 2002 |
| *r* | 57 | 166 |
| *w* | 9 | 54 |
| *y* | 3 | 3 |
| *hh* | 359 | 1373 |
| *iy* | 778 | 2357 |
| *ih* | 1126 | 3603 |
| *eh* | 232 | 767 |

| | | |
|---|---|---|
| *ey* | 140 | 492 |
| *ae* | 276 | 768 |
| *aa* | 90 | 224 |
| *aw* | 214 | 521 |
| *ay* | 383 | 1678 |
| *ah* | 309 | 1197 |
| *oy* | 0 | 0 |
| *ow* | 10 | 109 |
| *uh* | 5 | 13 |
| *uw* | 260 | 858 |
| *er* | 142 | 741 |

Table 2. *Phonetic accuracies for speaker dependent (SD) and speaker-independent (SI) systems*

| **Phoneme** | **Accuracy (%)** | | | |
|---|---|---|---|---|
| | **Without Warping** | | **With Warping** | |
| | ***SD*** | ***SI*** | ***SD*** | ***SI*** |
| *fr* | 44.7 | 28.0 | 47.7 | 50.6 |
| *m* | 45.3 | 16.1 | 55.6 | 41.6 |
| *n* | 40.9 | 18.3 | 42.5 | 38.3 |
| *ng* | 10.0 | 67.5 | 40.0 | 20.0 |
| *l* | 47.5 | 20.4 | 42.0 | 37.1 |
| *r* | 45.6 | 31.9 | 45.6 | 42.2 |
| *w* | 100 | 25.9 | 77.8 | 35.2 |
| *y* | 100 | 0.0 | 66.7 | 0.0 |
| *hh* | 35.4 | 23.8 | 22.8 | 33.0 |
| *iy* | 50.9 | 22.9 | 45.2 | 41.4 |
| *ih* | 48.2 | 21.2 | 45.9 | 41.2 |
| *eh* | 44.4 | 24.0 | 43.5 | 38.2 |
| *ey* | 52.1 | 19.1 | 44.3 | 31.9 |
| *ae* | 47.1 | 25.1 | 47.1 | 42.5 |
| *aa* | 64.4 | 53.1 | 64.4 | 73.4 |
| *aw* | 48.6 | 19.4 | 45.3 | 43.0 |
| *ay* | 48.6 | 28.2 | 44.9 | 43.1 |
| *ah* | 41.4 | 17.1 | 40.8 | 35.3 |
| *oy* | - | - | - | - |
| *ow* | 60.0 | 13.8 | 60.0 | 19.3 |
| *uh* | 60.0 | 7.7 | 20.0 | 84.6 |
| *uw* | 42.7 | 14.0 | 40.4 | 25.6 |
| *er* | 42.3 | 17.1 | 38.7 | 37.0 |
| ***Overall*** | **47.0** | **22.0** | **45.1** | **40.7** |

From Table 2 it can be seen that feature warping has very little effect on a speaker-dependent system, as expected. From Table 1, it can also be seen that the rate of occurrence of some phonemes higher than that of others, particularly semi-vowels and vowels. This is because better pitch estimates can be obtained from these phonemes than the others and only frames with pitch estimates were used in the experiments. Also, the accuracies for phonetic classes with very few test frames convey little or no useful information since they are easily affected by a few frames being misclassified (phonetic classes *//ng//*, *//w//*, *//y//*, *//oy//*, *//ow//*, *//uh//* can be safely ignored). Their low rates of occurrence also mean their contribution to the overall accuracy is negligible.

Unlike the speaker-dependent case, feature warping plays a very significant role in the speaker-independent system. This suggests that variations in the features between different speakers are quite large and much better modelling can be achieved when some sort of normalisation is used to reduce this variability. It is also interesting to note that the phoneme //aa// gives consistently high classification accuracies even for the case of a speaker-independent system without any speaker normalisation.

The Gaussian mixture models used for each emotional class in all of the abovementioned experiments were trained on data from all phonetic classes. It might be argued that better modelling may be achieved if a separate GMM was trained for every phonetic class for every emotion. During testing, since every test frame is associated to a particular phonetic class, maximum likelihood estimation is performed only over the five GMMs associated with the five emotions for that phonetic class. Such an experiment was performed for the speaker-independent case (there was insufficient training data to do this in a speaker-dependent manner) and the results are given below.

Table 3. *Phonetic accuracies for a speaker independent system using phoneme-specific GMMs*

| **Phoneme** | **Accuracy (%)** | |
|---|---|---|
| | **Without Warping** | **With Warping** |
| *fr* | 15.6 | 47.3 |
| *m* | 14.4 | 32.2 |
| *n* | 22.3 | 37.1 |
| *ng* | 40.0 | 7.5 |
| *l* | 24.3 | 43.4 |
| *r* | 34.9 | 52.4 |
| *w* | 37.0 | 44.4 |
| *y* | 0.0 | 0.0 |
| *hh* | 12.8 | 34.8 |
| *iy* | 28.8 | 42.7 |
| *ih* | 24.2 | 40.7 |
| *eh* | 23.7 | 34.0 |
| *ey* | 8.9 | 39.6 |
| *ae* | 21.1 | 40.0 |
| *aa* | 51.8 | 87.1 |
| *aw* | 13.8 | 41.8 |
| *ay* | 33.6 | 41.2 |
| *ah* | 11.0 | 36.0 |
| *oy* | - | - |
| *ow* | 15.6 | 31.2 |
| *uh* | 0.0 | 0.0 |
| *uw* | 31.5 | 25.5 |
| *er* | 18.5 | 42.4 |
| ***Overall*** | **22.6** | **40.0** |

Comparing the accuracies of the systems using phoneme-specific GMMs to those that use phoneme independent emotion models, the difference appears to be very small. This tends to suggest that the phoneme-specific models are very similar to the phoneme-independent models, indicating that for these features phonetic variability is very small and much less significant than speaker variability. However, none of the features used in the above experiments characterise the spectral content of speech, unlike the features used in almost all state-of-the-art phoneme recognition systems. This makes it hard to determine if the similarity of the phoneme-specific GMMs to the phoneme-independent GMMs is because of the lack of phoneme-specific information in the features or because the information being modelled by the emotion models are different from those modelled by phoneme recognisors. To clarify this, speaker independent emotion classification was performed with phoneme-independent and phoneme-specific GMMs using a MFCC based front end.

Table 4. *Phonetic Accuracies for a MFCC based Speaker Independent system (With Feature Warping)*

| **Phoneme** | **Accuracy (%)** | |
|---|---|---|
| | **Phoneme Independent GMMs** | **Phoneme Specific GMMs** |
| *fr* | 41.6 | 42.3 |
| *m* | 26.4 | 25.8 |
| *n* | 30.5 | 31.6 |
| *ng* | 17.5 | 15.0 |
| *l* | 34.1 | 37.8 |
| *r* | 31.9 | 43.4 |
| *w* | 25.9 | 50.0 |
| *y* | 100 | 100 |
| *hh* | 32.7 | 36.1 |
| *iy* | 37.2 | 37.3 |
| *ih* | 36.2 | 36.8 |
| *eh* | 37.7 | 39.2 |
| *ey* | 34.6 | 33.5 |
| *ae* | 42.7 | 26.9 |
| *aa* | 68.8 | 72.8 |
| *aw* | 47.6 | 40.1 |
| *ay* | 42.2 | 44.6 |
| *ah* | 40.6 | 37.0 |
| *oy* | - | - |
| *ow* | 22.0 | 72.5 |
| *uh* | 92.3 | 0.0 |
| *uw* | 33.6 | 40.0 |
| *er* | 34.8 | 33.5 |
| ***Overall*** | **36.9** | **37.2** |

From these accuracies it can be observed that once again there is very little difference between phoneme-specific and phoneme-independent emotion models. This leads us to believe that even when phoneme-specific information is present in the features, they are not modelled by the emotion models; lending further support to the observation that speaker variability is more significant problem to emotion modelling than phonetic variation.

## 4. Conclusions

The phoneme-specific emotion classification accuracies for a five-class problem reported in this paper allows us to conclude that differences in emotions are better conveyed by some phonemes than others, and that the accuracies of emotion models are affected to a larger extent by differences between speakers than they are by difference between phonemes. The consistently high classification accuracies for frames associated with certain phonemes, especially //aa//, indicates that emotion is conveyed by speech predominantly via these phonemes. Emotion classifiers that make a decision only based on frames from the phoneme //aa// would have a much higher accuracy than classifier that treats all phonemes equally, however, the low rate of occurrence of //aa// in speech might be a problem.

## 5. Acknowledgements

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

## 6. References

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

[2] Verceridis, D., Kotropoulus, C., and Pitas, I., "Automatic Emotional Speech Classification", in *Proc. IEEE ICASSP*, vol. 1, pp. I- 593-596, 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] Sethu, V., Ambikairajah, E., and Epps, J., "Group Delay Features for Emotion Detection," in *Proc. INTERSPEECH,* pp. 2273-2276, 2007.

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

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

[8] 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.

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

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

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

[12] Scwarz, P., Matejka, P., and Cernocky, J., "Heirarchical Structures of Neural networks for phoneme recognition", in *Proc. IEEE ICASSP,* vol. 1, pp. I- 325-328, 2008.

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