LATE INTEGRATION OF FEATURES FOR ACOUSTIC EMOTION RECOGNITION

Ailbhe Cullen, Naomi Harte

Dept. of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

ABSTRACT

It is widely accepted that the ability to understand emotion or affect from speech is central to the design of more natural human-computer interfaces. This paper explores the classification of natural emotional speech along four affective dimensions, using hidden Markov models (HMMs). A number of features are tested, some of which have never before been applied to emotion recognition. Finally, these different features are combined discriminatively to achieve a competitive performance on the AVEC 2011 affect classification task [1].

***Index Terms*—** Hidden Markov Model, Emotion recognition, Affect

1. INTRODUCTION

The desire to enable machines to understand the paralinguistic content of speech is driving a considerable research effort in the area of affective (i.e. emotional) computing [2, 3]. Early work in emotion recognition focused on the so-called basic emotion states: joy; sadness; anger; fear; disgust; surprise. However, this approach is limited in its ability to describe realistic human expression. An alternative is to rate speech on a number of affective dimensions, typically between two and four [3], each of which represents some emotion related quality. The four considered here are activation, valence, power, and expectation. Activation measures how passive or active the speaker is; valence measures the positivity or negativity of the speaker; power measures how in control the speaker appears to feel; and expectation measures the anticipation or surprise of the speaker.

A variety of features have been proposed for emotion recognition. These can be roughly grouped into four categories: prosodic; spectral; voice quality; and Teager Energy Operator (TEO) based. We consider a representative set from each of these categories. For comparison with previous work, we include a set of standard Mel-frequency cepstral coefficients (MFCCs), and a subset of the features from the recent Audio-Visual Emotion Challenge (AVEC 2011) [1], conducted at the 2011 International Conference on Affective Computing and Intelligent Interaction, which incorporates temporal, spectral, and pitch information. We also consider a set of TEO features and a set of voice quality features. All of these features are traditionally extracted over short (25-30ms) frame lengths. Since emotion varies more slowly than speech, we will also consider two sets of long-term spectro-temporal features [4, 5].

We use hidden Markov models (HMMs) to independently classify speech from the SEMAINE database, a corpus of natural emotional speech, along the four dimensions previously stated. One motivation for this is the established strength of the HMM framework within speech recognition, which emotion recognition seeks to supplement. Another is the potential of HMMs to capture the temporal evolution of emotion in speech at the highest level. For each dimension we implement five independent HMM classifiers, each using a different feature set. We then combine the outputs of the classifiers, using discriminative training to optimise the classifier weights. We compare our HMM approach with results obtained from a SVM classifier [1] and a multi-stage kNN/HMM system [6], both of which were tested on the same database.

This paper offers a number of novel contributions. Of the feature sets used, one [4] has never been applied to emotion recognition before, while another [5] has had only limited use. Some of the voice quality features we include have never before been used for emotion recognition. Finally, while some authors have attempted to combine audio and visual information, to the best of our knowledge there have been few attempts to fuse multiple acoustic classifiers. The closest study would be [7], which investigates early integration of acoustic features for emotion recognition, however the focus is on discrete emotions, and the authors do not explore late integration methods.

The rest of this paper is organised as follows. The emotional speech database will be discussed in Section 2. Section 3 will outline the feature extraction process and classifier structure. Results and discussion will be presented in Section 4. Finally, some conclusions will be given in Section 5.

2. DATABASE

The database used is the audio portion of the SEMAINE database of emotionally coloured character interactions [8]. The database consists of conversations held between an operator and a user. The operator adopts the role of one of four characters, and by acting emotionally attempts to induce nat-

This work is funded by the Irish Research Council under the Embark postgraduate scholarship scheme. Many thanks to John Kane for the voice quality features.

Table 1. Full list of features used. *extracted using Aparat [9] †extracted using openSMILE [10]

Voice Quality	OQ1, OQ2, OQa, QOQ, AQ, NAQ, ClQ, SQ1, SQ2, H1-H2, HRF	[9]*
	Peak Slope (Wavelet, Glottal)	[11]
	RPP, H2-H1	[12]
	MDQ	[13]
Prosodic	Intensity, Zero Crossing Rate, Energy in bands 250 - 650 Hz, 1 - 4 kHz, 25%, 50%, 75%, 90% spectral roll-off, Flux, Entropy, Variance, Skewness, Kurtosis F0, Probability of voicing, jitter, delta jitter, shimmer, log(HNR)	[1]†
MFCC	MFCC 1-12	
TEO	TEO-FM-Var, TEO-Auto-Env, TEO-CB-Auto-Env	[14]

ural emotional responses in the user. This has resulted in a rich database of over 12 hours of natural emotional speech.

The SEMAINE database was recently used as the challenge data for AVEC 2011 challenge [1]. The task for this challenge was to classify words along four affective dimensions, using either audio or video information, or both. Only a binary classification (High/Low) was performed, and at word level. Since we are concerned only with audio in this study, we will use the audio sub-challenge as a reference with which to compare our results.

3. EXPERIMENTAL SETUP

A full list of the short term features used in this study and relevant references is given in Table 1. Unless otherwise stated features are extracted over 25ms frames (hamming windowed), spaced 10ms apart. The MFCC, TEO, and AVEC features have been discussed extensively in previous literature [1, 3, 14], thus we will focus the following discussion on the voice quality and long term modulation spectrum (LTMS) features. The dimensionality of all feature sets, except for the MFCCs, is reduced via principal component analysis (PCA), and first derivatives of all features are computed over a three frame window.

3.1. Voice Quality Features

The majority of existing voice quality features are parametrisations of the glottal waveform. Thus we first record 11 time and frequency measurements from the glottal waveform as in [15]. It is known that certain irregular phonation types (strong determinants of voice quality) cause difficulties for glottal source estimation [16]. Therefore, we include a number of new voice quality features which do not rely on glottal source estimation, and have not previously been used for emotion recognition. The residual peak prominence and difference in 1st and 2nd harmonics (H2-H1) are calculated from the linear prediction (LP) residual and have been designed to classify a particular type of irregular phonation, namely creak [12]. The peak slope and maximum dispersion quotient (MDQ) are estimated from a wavelet decomposition [11, 13] and are designed to capture breathy or tense voice.

3.2. Long Term Features

Long term modulation spectrum (LTMS) features are also extracted. Two different approaches are explored [4, 5]. Both involve critical band filtering the speech signal, but differ in their subsequent modelling of human auditory perception. The first set (which we will refer to as the LTMS(A) set [4]) includes an adaptive compression loop which accentuates sudden changes in the sub-band envelope and suppresses slow changes. These features are extracted over frames of 600-900 ms. The second set (LTMS(B) [5]),which has previously been used for emotion recognition, incorporates a number of measures of the energy distribution of a spectro-temporal representation of each sub-band, extracted over 250 ms frames.

In order to understand the importance of the frame-length, both LTMS feature sets were extracted over frame lengths between 250 ms and 900 ms, with a 10 ms delay between frames.

3.3. HMM Classifier

We use a left/right HMM classifier, in which HMMs were trained using frame level features. As mentioned above, the optimum model for emotion recognition is still an open question. Furthermore, it is possible that the different dimensions may be best modelled by different HMMs. We tested a range of HMM models, containing between one and five states, and with up to fifteen Gaussian mixtures per state. The HMMs were implemented using the Hidden Markov Model Toolkit (HTK), and were trained using word level labels. For each dimension, two HMMs were trained, corresponding to the High and Low binary labels.

A noted drawback of the HMM is its inability to model supra-segmental information [3]. The HMM classifier models the evolution of emotion within words, but does not account for long term dependencies between words. Emotion varies slowly, so rapid switching between high and low affective states is unlikely. Thus a post processing stage is included in which a median filter is applied over a seven word window to de-noise the output labels.

3.4. Decision Fusion

In our final experiment we attempt to exploit the strengths of each individual classifier by combining the results of each individual system. For each classifier, k, and test word x_n, we can define L_{njk} as the log likelihood that the word x_n belongs to the class C_j. The overall log likelihood of the class C_j given the results obtained from the five classifiers for the n^{th} word is approximated as follows

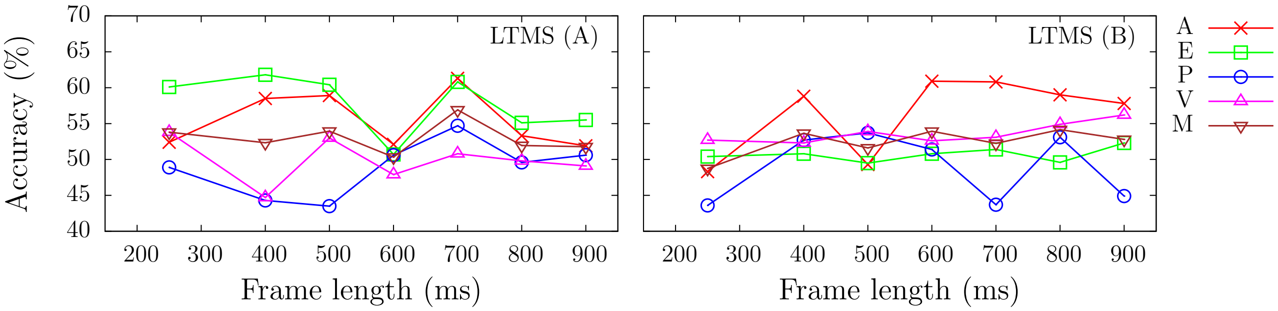

Fig. 1. Classification performance of Long Term Modulation Spectrum (LTMS) features extracted over a range of frame lengths on four dimensions: A(ctivation), E(xpectation), P(ower), V(alence), and mean performance (M) over all four.

$$L_{nj} = \sum_{k=1}^{K} w_k L_{njk} \tag{1}$$

where $\{w_k\}$ are the classifier weights, and $K = 5$ is the number of independent classifiers to be combined.

The decision rule is thus

$$C(x_n) = C_i \quad \text{if} \quad L_{ni} = \max_j L_{nj} \tag{2}$$

We explore three approaches to choosing the classifier weights, $\{w_k\}$: assigning equal weights; via Minimum classification Error (MCE) training; and via Maximum Mutual Information (MMI) training [17].

4. RESULTS

We compare our results with two references. The first is the baseline given by the organisers of the AVEC 2011 audio subchallenge, which was obtained using a SVM classifier [1]. The second is performance of a hybrid kNN/HMM system proposed by Meng et al. [6], which won the AVEC 2011 audio sub-challenge. All HMMs are trained on the AVEC training set, and unless otherwise stated are tested on the AVEC test set. The performance measure is the weighted word-level accuracy (WA) as used in the AVEC 2011 challenge [1].

4.1. Dynamic Modelling of Emotion

We explore the incorporation of temporal information in two ways: by varying the LTMS feature length, and the number of states and mixtures in the HMM model. The HMM classifier models short term (within word) evolution of emotion, but does not capture long term trends. The smallest frame length considered for the LTMS features (250 ms) is close to the average word length (263 ms), thus the LTMS features may capture higher level patterns in emotion.

Figure 1 shows the change in performance of the two LTMS feature sets on the AVEC development partition as the frame length is varied. Overall the performance on activation and power is roughly comparable for both LTMS sets. Expectation is better captured by the LTMS(A) set while valence favours the LTMS(B) set. For both LTMS classifiers valence is the least affected by the frame length. The best performance on activation is achieved between 600 and 700 ms for both features. The classification of expectation using the LTMS(A) set is generally better for shorter frame lengths (less than 600 ms) while using the LTMS(B) set the performance varies little with frame length. The best performance for the classification of power with the LTMS(B) set is achieved with frame lengths between 400 ms and 600 ms, while using the LTMS(A) set longer frame lengths are preferable.

By calculating features over longer frames, we introduce a degree of memory which begins to compensate for the aforementioned inability of the HMM to model supra-segmental information. Given that the performance on activation and power is maximised at around 600 ms, while expectation and valence appear to be unaffected by the changing frame length, or to favour shorter frames, we conclude that this long term memory is particularly useful for activation and power.

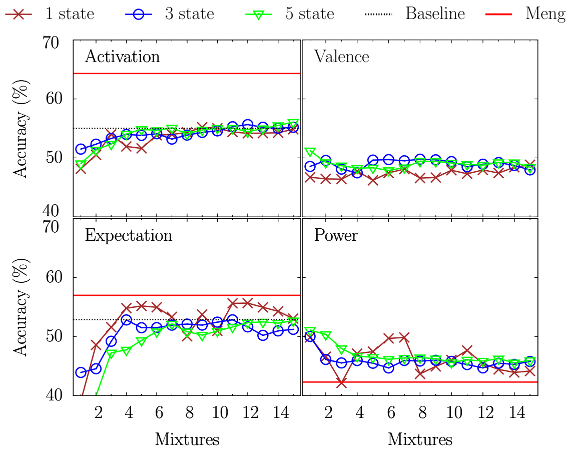

Fig. 2. Performance of HMM classifiers with varying numbers of HMM states and mixtures, using the AVEC-r feature set, compared with the AVEC baselines and the performance of the Meng classifier system.

Figure 2 compares the performance of a range of HMM topologies, using the AVEC-r features set, on the AVEC test partition. For activation, once the number of mixtures is increased to 10 there is no advantage achieved by increasing the number of states. Similarly for power, the performance of the single state HMM with 6 or 7 mixtures is close to the performance of the three and five state HMMs with just one mixture. The best performance on expectation is achieved

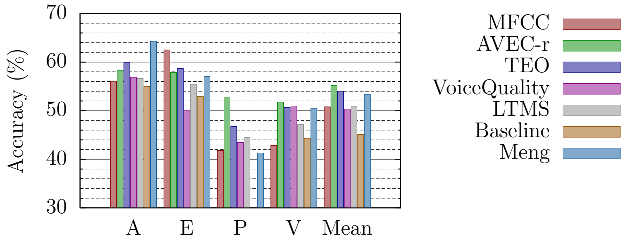

Fig. 3. Comparison of feature sets for HMM affective classification. The LTMS set used here is LTMS(B) with a frame length of 600 ms

with the single state HMM and the performance consistently falls as the number of states increases. This suggests that for activation, expectation, and power, provided we have sufficient mixtures to model the diversity of the data, there is no further benefit to be obtained by modelling the temporal evolution at a state level. The only dimension which clearly benefits from the dynamic modelling ability of the HMM is valence. There is a small but consistent advantage to using a 3 state model with 6 to 10 mixtures on this dimension.

The sharp decrease in performance of the single state HMM on power when the number of mixtures is increased from 7 to 8, and the steady decline in performance of the 3 and 5 state models, also suggests that there is insufficient variation in our data to train more complex models. The high and low labels for both power and valence are unevenly distributed in the training partition. The relative scarcity of examples of Low may explain why neither dimension appears to benefit from a large number of Gaussian mixtures.

4.2. Comparison of Feature Sets

Figure 3 compares the performance of the five feature sets with the challenge baseline and the performance of the Meng et al. [6] classifier. The best overall performance is 54%, given by the voice quality features, which just surpasses the 53.3% of Meng et al.. This is largely due to the significant improvement achieved on the power dimension. The voice quality features also outperform Meng et al. on valence. None of the feature sets tested improve upon the classification of activation achieved by Meng et al. However, MFCC feature set performs quite well on expectation.

4.3. Performance of ensembles

An important consideration when combining classifiers is the diversity of the classifier set. In general, the more diverse the classifier set, the better the results of the combined classifier. Therefore, before reporting the performance of the ensemble, we briefly investigate the diversity.

Figure 4 shows the frequency with which at least n of the classifiers correctly classify a word from the development set. The $n = 3$ column corresponds to the performance of a majority vote, and varies between 40% and 60% depending on the dimension in question. We also measured diversity using three measures described in [18]. These are reported in Table 2. A diverse classifier set should have Entropy and KW variance close to 1 and inter-agreement κ close to zero. The entropy and inter-agreement scores from Table 2 suggest that the classifier set is quite diverse. Therefore we would expect a reasonable improvement in accuracy from the combined classifier. The small KW variances may be due to the fact that we consider a two class problem, so we cannot realistically expect a large variance.

Table 2. Entropy (E), Kohavi-Wolpert variance (KW), and inter-agreement (κ), measures for the four classifier ensembles

	E	KW	κ
A	0.523	0.140	0.293
E	0.734	0.190	0.049
P	0.696	0.182	0.086
V	0.710	0.184	0.077

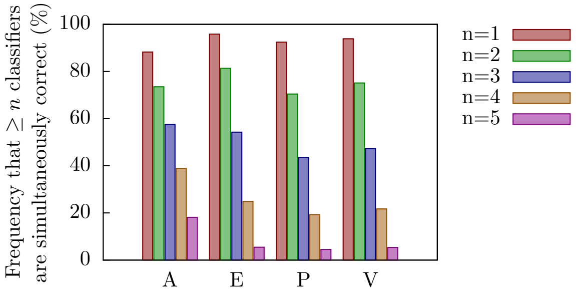

Fig. 4. Frequency that at least n classifiers correctly classify words from the development set

Table 3 outlines the result of fusing the individual classifier outputs via a weighted sum of log likelihoods using equal classifier weights (Average), using MCE trained weights, and using MMI trained weights. The performance of the overall best individual feature set, the voice quality (VQ) set, and the performance reported by Meng et al. [6], are also given for comparison. Overall, the average log likelihood performs best. This is counter intuitive. We would expect the MCE or MMI trained weights to boost classifiers which perform well and suppress classifiers which perform poorly, thus improving the overall results even further than the equi-weighted sum. However, the discriminative training cannot take into account how well a given classifier will generalise to unseen data. The test set is speaker independent, and emotion and affect are person specific, particularly in natural speech. For example, the best classification on valence in the test partition is given by the voice quality features. These same features perform worst on valence in the training set.

Overall, the combination of multiple classifiers increases performance by 1.9% over that of the voice quality set, and by 2.6% over that acheived by Meng et al. [6]. Given the diversity results in Table 2 and the information of Figure 4, better performance is possible. Clearly we need a more sophisticated method of learning when to trust each classifier.

It is worth noting that the results in Table 3 are not directly comparable to Figure 2 since the HMM topology is not fully optimised for this experiment. However, we would expect the improvements from classifier combination to generalise to other HMM structures.

Table 3. Summary of performances achieved per dimension by ensemble classifiers.

	Meng	VQ	Average	MCE	MMI
A	64.3	58.2	59.5	59.4	59.6
E	57.0	55.1	56.4	56.4	56.2
P	41.3	52.1	51.5	48.9	52.2
V	50.5	50.8	56.4	43.8	44.5
Mean	53.3	54.0	55.9	52.1	53.1

5. CONCLUSION

From the above analysis is it clear that some degree of temporal modelling is required for affective classification. Valence which is traditionally considered captured modelled by acoustic features benefits the most from the use of HMMs over GMMs. This suggests that the short-term evolution of acoustic features is important for the identification of this dimension, while long-term trends, introduced by the LTMS features are more useful for activation and expectation. This concept of including multi-level temporal information warrants further investigation.

Of the five feature sets explored, the voice quality features performed the best on average. This is most likely because the irregular phonation types which humans use to express emotion often cause difficulties for automated feature analysis. The voice quality set contains features are specifically designed not only to be robust to these irregular phonation types, but also to identify them.

Regarding the reproducibility of these results, a study by Ntalampiras et al. [7] shows a benefit to both temporal modelling and feature fusion for the recognition of discrete emotions. However when we carried out experiments on the FAU-AEC database [19] no similar trends were found in feature performances or HMM structure. This may be because acoustic emotion expression is age, language, or culture dependent (SEMAINE consists of English-speaking adults while FAU-AEC contains German-speaking children).

Given the high classifier diversity, the performance of the ensemble classifier is disappointing. Work is ongoing to explore more involved fusion techniques. Given the complexity of emotions in speech, it is reasonable to expect that more complex or involved approaches to classifier combination are required.

6. REFERENCES

[1] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, *AVEC 2011The First International Audio/Visual Emotion Challenge*, vol. 6975 of *LNCS*, pp. 415–424, Springer, 2011.

[2] R. Picard, *Affective Computing*, MIT Press, Cambridge, Massachussets, 1997.

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Sp. Comm.*, vol. 53, no. 910, pp. 1062–1087, 2011.

[4] S. Ganapathy, S. Thomas, and H. Hermansky, "Static and dynamic modulation spectrum for speech recognition," in *Interspeech*, 2009, pp. 2823–2826.

[5] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Sp. Comm.*, vol. 53, no. 5, pp. 768–785, 2011.

[6] H. Meng and N. Bianchi-Berthouze, *Naturalistic Affective Expression Classification by a Multi-stage Approach Based on Hidden Markov Models*, vol. 6975 of *LNCS*, pp. 378–387, Springer, 2011.

[7] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 116–125, 2012.

[8] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[9] M. Airas, "Tkk aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[10] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010, pp. 1459–1462.

[11] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Interspeech*, 2011, pp. 177–180.

[12] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, no. 4, 2013.

[13] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 21, no. 6, pp. 1170–1179, 2013.

[14] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 3, pp. 201–216, 2001.

[15] R. Sun and E. Moore, *Investigating Glottal Parameters and Teager Energy Operators in Emotion Recognition*, vol. 6975 of *LNCS*, pp. 425–434, Springer, 2011.

[16] J. P. Cabral, J. Kane, C. Gobl, and J. Carson-Berndsen, "Evaluation of glottal epoch detection algorithms on different voice types," in *Interspeech*, 2011.

[17] V. Valtchev, *Discriminative Methods in HMM-based Speech Recognition*, Ph.D. thesis, 1995.

[18] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[19] Stefan Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, 2009.