INTERSPEECH 2009 BRIGHTON

Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge

Iker Luengo, Eva Navas, Inmaculada Hernáez

Department of Electronics and Telecommunication, University of the Basque Country, Spain
{iker.luengo, eva.navas, inma.hernaez}@ehu.es

Abstract

This paper describes the system presented at the Interspeech 2009 Emotion Challenge. It relies on both spectral and prosodic features in order to automatically detect the emotional state of the speaker. As both kinds of features have very different characteristics, they are treated separately, creating two sub-classifiers, one using the spectral features and the other one using the prosodic ones. The results of these two classifiers are then combined with a fusion system based on Support Vector Machines.

Index Terms: expressive speech, emotion, feature selection

1. Introduction

The field of automatic emotion recognition from speech has gained interest among the speech processing research groups in the last couple of years. This interest increase is largely related to a fast development of the technology, which has allowed the creation of new machines with more natural speech-based interfaces, like pet-robots or avatars. Nevertheless, although there is a large number of published works in this field [1, 2, 3], the results are far from being comparable to the ones obtained by similar fields as the automatic speech recognition and the speaker identification. The reason for these (relatively) poor results is the great complexity of the emotions. Human emotions are expressed as a complex combination of small physiological responses (see [4] for a good review) that makes the problem hard to analyse by automatic means. As a result, it seems reasonable to think that a single type of feature will not be sufficient to achieve a good recognition accuracy, and that many features of different nature should be used for the task. This has been the approach in the last years [5, 1], nevertheless, the problem is still far to be solved.

The Interspeech 2009 Emotion Challenge [6] has been created in order to provide a standardised environment for testing the performance of different emotion recognition systems and to allow direct comparison among results. This paper describes the system presented to the challenge and the obtained results. With only the voice as a clue for such a complex task, the system is designed to retain as much information about the emotion as possible. It uses both spectral envelope (vocal tract information) and prosodic (glottal source information) features. Spectral and prosodic features are modelled independently and a classifier fusion technique is used to combine the results of both sub-systems.

2. Spectral features

Spectral information was parametrised with Mel-scale short-time log-frequency power coefficients (LFPC) [7]. Specifically 18 LFPC were used together with first and second derivatives, giving a grand total of 54 features. Features were calculated every 10 milliseconds, using Hamming windowing of 25 milliseconds. Mean normalisation was applied to every segment in order to eliminate the effect of the distance to the microphone. No other type of normalisation was applied.

3. Prosodic features

Long-term statistics of different prosodic values were used for the prosodic characterisation of the emotions. A Voice Activity Detector (VAD) algorithm based on [8] was used in order to detect the pauses in the recordings, and a feature vector was extracted for every between-pauses segment, i.e., the time between consecutive pauses was taken as the integration time for the calculation of the statistics.

All prosodic features are derived from intonation and energy values extracted for every frame as well as from an automatic vowel detector.

3.1. Intonation features

Intonation values together with voiced-unvoiced labels were estimated every 10 milliseconds with the algorithm described in [9]. Then first and second derivatives of the intonation curves were calculated. For each of these three curves, the following statistics were computed for each between-pauses segment in order to retain the intonation-related prosodic information: mean ($E\{\cdot\}$), variance ($\sigma^2\{\cdot\}$), minimum ($min\{\cdot\}$), range ($R\{\cdot\}$), skewness ($Sk\{\cdot\}$) and kurtosis ($Kr\{\cdot\}$).

Only frames detected as voiced were used for the estimation of the statistics. This gives 18 intonation-related features.

3.2. Power features

Power values were calculated every 10 milliseconds. In order to remove the effect of the distance to the microphone, the mean power value of each recording was estimated (discarding silence segments as detected by the VAD algorithm) and this value was subtracted from every sample.

Similarly to the intonation-related features first and second derivatives were calculated for the power curve and the same statistics were computed, giving another 18 power-related features. Note that mean power normalisation was performed over the whole recording, and that the *mean power* feature is calculated over a between-pauses segment, so this feature does not have to be zero.

3.3. Rhythm features

Rhythm features considered in the system are the mean duration of vowels ($E\{Vdur\}$) and the variance of vowel duration

$(\sigma^2\{Vdur\})$.

Vowels in the recordings were automatically detected using a simple vowel detector based on hidden Markov models (HMM). Each vowel was modelled by a different model, while groups of consonants shared a single model. These groups of consonants were automatically created by a clustering algorithm based on acoustic similarity among them. Not surprisingly, the resulting clusters were easily recognisable. For example, there is a cluster for nasals, another one for fricatives and so on. Both vowels and consonants used 1024 mixture three-state left-to-right models.

The accuracy of this simple vowel detector was tested over the training set, obtaining an 80% accuracy.

3.4. Regression features

Some researchers in emotion identification in speech consider not only global statistics of intonation and power over the whole sentence, but also the tendencies of these curves on shorter syllable-type segments [10]. As we already used an automatic vowel detector, we decided to use these detected segments. For each vowel, the linear regression of the power and intonation curves within the vowel boundaries was estimated, and new features were extracted from the mean value and slope of the regression:

- Mean of absolute values of intonation slopes in vowels $(E\{F0sl\})$
- Variance of absolute values of intonation slopes in vowels $(\sigma^2\{F0sl\})$
- Maximum of absolute values of intonation slopes in vowels $(max\{F0sl\})$
- Mean of absolute values of power slopes in vowels $(E\{POWsl\})$
- Variance of absolute values of power slopes in vowels $(\sigma^2\{POWsl\})$
- Maximum of absolute values of power slopes in vowels $(max\{POWsl\})$

3.5. Voice quality features

Jitter and shimmer values for each segment were used as a measure of the voice quality, as some emotions may be characterised by these features. Jitter (Jit) stands for micro-variations of the intonation curve, whereas shimmer (Shm) is the measure of micro-variations in the power curve.

3.6. Sentence-end features

Prosodic values related to the end of the sentence (e.g. pitch and energy rise/fall or syllable lengthening/shortening) may give additional clues, as emotional speech often has a special effect on sentence endings. Therefore some features related to the last vowel detected in the segment were also considered, namely:

- Normalised and unnormalised slope of intonation in last vowel $(LvF0sl) - (NLvF0sl)$
- Normalised and unnormalised central value of intonation in last vowel $(LvF0cn) - (NLvF0cn)$
- Normalised and unnormalised slope of power in last vowel $(LvPOWsl) - (NLvPOWsl)$
- Normalised and unnormalised central value of power in last vowel $(LvPOWcn) - (NLvPOWcn)$
- Normalised and unnormalised duration of last vowel $(Lvdur) - (NLvdur)$

5-class problem		2-class problem	
Rank	*Feature*	*Rank*	*Feature*
1	$R\{\partial POW\}$	1	$R\{\partial POW\}$
2	$E\{Vdur\}$	2	$min\{\partial^2 POW\}$
3	$E\{F0\}$	3	$E\{F0\}$
4	$Kr\{F0\}$.	4	$E\{POW\}$
5	$E\{POW\}$	5	$Lvdur$.
6	$NLvF0sl$	6	$NLvF0sl$
7	$min\{\partial^2 POW\}$	7	$Kr\{\partial^2 F0\}$
8	$min\{\partial POW\}$	8	$min\{\partial POW\}$
9	$\sigma^2\{Vdur\}$	9	$Sk\{\partial^2 POW\}$
10	$Lvdur$	10	Shm

Table 1: 10 best prosodic features as ranked by the feature ranking algorithm.

The normalised value of a feature stands for the value of the feature divided by its mean value for all vowels in the parametrised segment. Altogether there are 10 features related to the end of the segment.

4. Feature ranking

Some of the considered prosodic features may not be relevant for the classification of emotions in the speech. Some of them may not give any information about the emotion, and others may be correlated among them, therefore, not giving new information and being redundant. Using irrelevant or redundant features may decrease the accuracy of a classifier, due to the confusion that they add to the system. Therefore a feature selection criterion is needed in order to select those features that really give discriminant information to the system.

In this work a forward selection algorithm [11] based on inter-class and intra-class distances was used. Let S_W and S_B be the within-class scatter matrix and between-class scatter matrix respectively, defined as follows:

$$S_W = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(y_i^j - \bar{y}_i)(y_i^j - \bar{y}_i)^T \quad (1)$$

$$S_B = \frac{1}{N}\sum_{i=1}^{M} N_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \quad (2)$$

where $N = \sum_{i=1}^{M} N_i$ is the total number of training samples, y_i^j are the samples of class i, M is the number of classes and $\bar{y}_i$ and $\bar{y}$ are the class mean and global mean respectively:

$$\bar{y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_i^j \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} y_i^j \quad (3)$$

The criterion for maximising the inter-class separation and minimising the intra-class separation at the same time is the J_1 criterion described in [12]. This criterion is commonly used for LDA in the literature, and it is equivalent to the well-known Fisher discrimination criterion $J_{Fisher} = \frac{|S_B|}{|S_W|}$. So the ranking algorithm uses the criterion:

$$J_1 = \mathrm{tr}\left(S_W^{-1} \cdot S_B\right) \quad (4)$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

Table 1 shows the 10 best features as ranked by this algorithm for both the 5 class and 2 class problem.

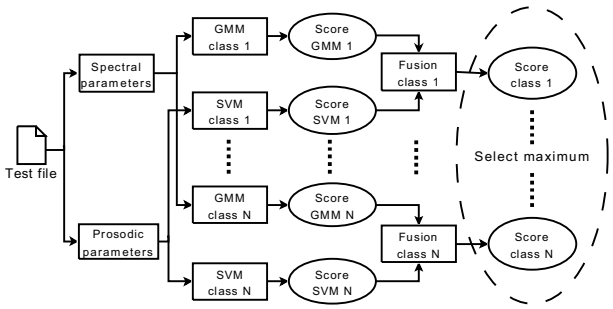

Figure 1: Schematic view of the emotion identification system.

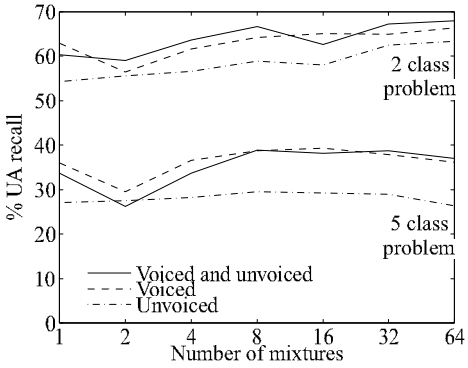

Figure 2: Results of development tests for spectral features.

5. Classifiers

5.1. System overview

Each feature set (spectral and prosodic) has very different properties, the most important of them being the number of available samples for both training and testing. While spectral features are extracted every 10 milliseconds, a single prosodic feature vector is extracted for each between-pauses segment. Taking into account that most of the recordings have a single speech segment (i.e., without middle pauses), most of the recordings are parametrised with a single prosodic vector. This gives rather few data for prosodic model training compared with the quantity of data available for spectral model training. Thus, different modelling approaches were used for each parametrisation.

Figure 1 shows a general view of the emotion identification system and the classifiers used. Spectral features are used to train a different Gaussian mixture model (GMM) for each emotion. In a similar way, a Support Vector Machine (SVM) [13] is trained for each emotion using prosodic features. During test phase, the scores obtained from the GMM and SVM models for one emotion are combined in order to obtain a single score per emotion. Finally, the emotion corresponding to the highest score is selected as the final class.

In order to train the fusion system as well as to choose the appropriate meta-parameter values for the prosodic and spectral models, some development tests are needed. To get these the speakers available in the training set were randomly distributed into five blocks, each one with five speakers (speaker 20[1] was discarded in order to have blocks with the same number of speakers). Over these blocks a leave-one-out loop was applied, where five different systems were trained using four of the blocks and tested on the remaining one. After using the results of these tests for development, the final system was trained using all speakers.

5.2. Spectral modelling: GMM

For the LFPC features GMM were used, as enough features were extracted from each recording as to train robust models. During development tests, 1 to 128 mixture models were evaluated in order to select the appropriate mixture number for the final system. The results of these tests are shown in Figure 2 for both the 5-class and 2-class problem. The performance reaches a maximum around 16–32 mixtures for the 5-class problem and around 32–64 mixtures for the 2-class problem. Therefore the use of 32 Gaussian mixtures was decided for the final system.

[1]Speaker 20 was selected because, although speaker 22 has fewer training samples (168 against 181), most of the samples from speaker 20 are in neutral style, which is already over-represented.

5.3. Prosodic modelling: SVM

In the prosodic parametrisation a single feature vector is extracted for each between-pauses segment. As most of the recordings have no middle pause, most of them are parametrised with a single prosodic vector. Specifically, for the training set we found that:

- 8644 recordings are parametrised with a single vector
- 1099 recordings are parametrised with two vectors
- 216 recordings are parametrised with more vectors

As there is not so much data for model training, a SVM classifier with RBF kernel was selected for prosodic modelling. SVM's have proved to be very accurate and have high generalisation capability when few training data is available [14]. A one-against-all approach was taken for the multiclass (5-class) problem, i.e. five classifiers were trained trying to separate each class from the rest. When more than one speech segment was detected, the final score was calculated as the product of the scores for each segment.

Prior probabilities of each class should not be taken into account, as the final performance measure will be unweighted recall. To compensate the imbalance in the number of training samples for each class, a different missclassification cost value (a SVM meta-parameter usually denoted by C) was used during training for each class. Classes with more training examples were assigned lower cost values, reducing their influence, and thus partly compensating their higher prior probability.

In order to select the number of features that should be used in the final system, several prosodic classifiers were trained using sets of one, two... up to 56 features, following the ranking obtained in section 4. Figure 3 show the system performance against the number of features, as estimated during development tests. In both the 5 class and the 2 class problems the performance stabilised after 15 features, but there is a local maximum around 10. For this reason 10 prosodic features were used in the final system. These 10 features are shown in Table 1.

5.4. Classifier fusion

The scores obtained from the spectral and prosodic systems were combined using a SVM classifier. This score fusion was done independently for each emotion, as shown in Figure 1. This fusion schema, considering each emotion independently, has given us better results that combining all scores into a single SVM fusion system.

The training of these fusion SVM's was performed using the results of the development tests described in section 5.1.

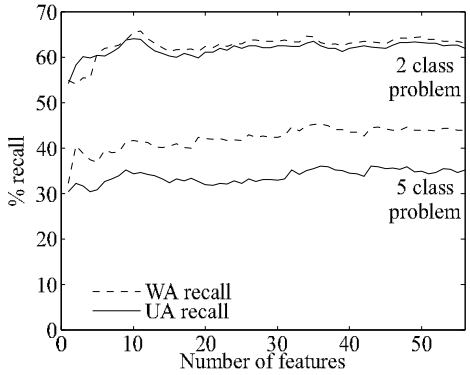

Figure 3: Results of development tests for prosodic features.

	5-class	2-class
Prosodic (10 feat.)	35.39%	60.09%
Spectral (32 mixt.)	38.73%	61.46%
Fusion	41.38%	67.19%

Table 2: Summary of the UA recall for the tested solutions.

For each of the development test recordings, the spectral and prosodic scores were calculated for each emotion. Then, a fusion SVM was trained for each emotion. When training emotion X, all development test recordings belonging to that emotion were labelled '1' while the rest of recordings were labelled '0', so that the fusion classifier would learn to separate the correct scores of that emotion from the wrong ones.

6. Results and conclusions

Table 2 summarises the results obtained with this system in the challenge, in terms of UA recall. Not only the final system was evaluated, but also the intermediate prosodic and spectral systems. Comparing the results for these two sub-systems alone with the ones obtained in the development tests it can be seen that the system performance is very close to the expected one, although it is not so accurate in the 2-class problem. In both cases results for spectral features are slightly better than for prosodic ones. Nevertheless, when combining both sub-systems results do improve a 7% in the 5-class problem and a 10% in the 2-class one.

For a deeper analysis of the results, Tables 3 and 4 show the confusion matrix for the final (fused) system in both problems. Recordings corresponding to the **R**est class are the most difficult to detect, probably because it is an heterogeneous class and each recording may have characterisitcs related to other emotions. **A**nger and **E**mphatic have a high confusion between them, which is not surprising as both are emotions with a high activation degree. Another confusion-pair is formed by

	A	E	N	P	R	Sent.
A	67.76	15.88	7.53	4.26	4.58	611
E	28.38	45.56	18.83	2.98	4.24	1508
N	24.21	20.25	43.52	7.87	4.15	5377
P	13.49	5.12	36.28	40.93	4.19	215
R	31.5	11.72	28.02	19.6	9.16	546

Table 3: Confusion matrix for the whole system in the 5-class problem. (rows: reference; columns: hypothesis; values in %)

	NEG	IDL	Sent.
NEG	76.96	23.04	2465
IDL	42.58	57.42	5792

Table 4: Confusion matrix for the whole system in the 2-class problem. (rows: reference; columns: hypothesis; values in %)

Emphatic and **N**eutral. Even though we tried to compensate the *a priori* probabilities several recordings are misclassified as **N**eutral, the class with most training examples. Many are also misclassified as **A**nger although it has much fewer examples.

Although the results are not impressive, it has to be taken into account that the task in the challenge was a hard one, where no prototypical overacted emotions were used, but spontaneous natural ones.

7. Acknowledgements

This work was partially supported by the Avivavoz project, MEC (TEC2006-13694-C03-02/TCM) and the ANHITZ program of the Basque Government (IE06/185).

8. References

[1] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, Oct. 2008.

[2] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, Oct. 2007.

[3] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[4] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. Journal of Human-Computer Studies*, vol. 59, pp. 157–183, 2003.

[5] T. Vogt and E. Andre, "Improving automatic emotion recognition from speech via gender differentiation," in *Fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006.

[6] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, Brighton, UK, Sep. 2009.

[7] T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, June 2003.

[8] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long term speech information," *Speech Communication*, vol. 42, pp. 271–287, Apr. 2004.

[9] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *ICASSP*, Honolulu, USA, Apr. 2007, pp. 1057–1060.

[10] F. Ringeval and M. Chetouani, "Exploiting a vowel based approach for acted emotion recognition," *Lecture Notes on Computer Science*, vol. 5042, pp. 243–254, 2008.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, A. Press, Ed., Boston, USA, 1990.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[14] V. D. Sanchez A., "Advanced support vector machines and kernel methods," *Neurocomputing*, vol. 55, pp. 5–20, Sep. 2003.