Feature analysis and evaluation for automatic emotion identification in speech

Iker Luengo, Eva Navas

University of the Basque Country
Alda. Urquijo s/n 48013 Bilbao (SPAIN)
iker.luengo@ehu.es, eva.navas@ehu.es

Abstract

This PhD dissertation was written by Iker Luengo under the advice of Dr. Eva Navas. It was successfully defended in the University of the Basque Country the 1^{st} of June, 2010. The doctoral thesis committee was composed by Prof. José B. Mariño as president, Prof. Inmaculada Hernáez as secretary, Prof. Carmen García, Dr. Valentín Cardeñoso and Dr. Laura Docío.

Index Terms: emotion identification, parameterisation, feature selection

1. Introduction

Features extracted from the speech signal have a great effect on the reliability of an emotion identification system. Depending on these features, the system will have a certain capability to distinguish emotions, and will be able to deal with speakers not seen during the training. Many works in the field of emotion recognition are aimed to find the most appropriate parameterisation, yet there is no clear agreement on which feature set is best.

Going over the literature shows that each research group uses a different parameter set. Features extracted from the spectral envelope, prosodic characteristics, glottal flow or even the linguistic content are used indiscriminately, in an attempt to detect those that really are useful. The approach that is most often used is to define a large parameter set and feed it to an automatic feature selection algorithm that heuristically selects the most discriminant ones. i.e., let the data tell you what is relevant and what not [1, 2]. Unfortunately, this features selection step is usually seen as an unavoidable nuisance needed in order to maximise the accuracy, not as a useful tool for the enlightening of the relation between the features and the emotions. Only few papers show the results of this selection, and almost none discuss them, so it is not possible to know if prosody, spectral envelope or voice quality features were the ones that really provided information to the system.

It is widely accepted that prosodic features carry most of the emotional information in the speech. For many years automatic identification systems have used prosody almost exclusively, leaving spectral characteristics in a second term [3, 4]. There are several reasons behind this idea. On the one hand, the theories describing the physiological changes caused by an emotional state [5, 6] focus mostly on changes in the air pressure and vocal cord tension. On the other hand, many studies directly compare speech signals with different emotional content in order to find measurable differences [7, 3]. These studies usually conclude that, taken individually, the differences from one emotion to another are larger for prosodic features than for spectral ones. But the behaviour of the complete set of features is not analysed, so the complete set of spectral characteristics may provide more information than the complete set of prosodic features. In fact, it is nowadays very usual to find papers successfully using spectral features [8, 9, 10] or voice quality characteristics [11, 12], thus proving that these kind of parameters are also important in the classification of emotions.

Furthermore, no systematic study of the effectiveness of each parameterisation has been performed. Such a study could identify which is the most appropriate parameterisation for the automatic identification of emotions. There are some works in the literature that analyse the behaviour of different feature sets, but they do not provide a complete view of the problem. Many of these studies treat each feature individually, which may provide conclusions that are not generalisable when they are combined with others. Other works provide experimental results with complete parameterisations, but not separately, so it is not possible to deduce whether the combination of features was really better than the features (or a feature subset) alone.

Finally most of the works in the literature are not comparable among them, as they use different speech databases, different methodologies or different number of emotions. As a result, it is impossible to build a complete view of the properties of different parameterisations comparing the results of such works.

2. Objectives

The work presented in this dissertation attempts to fill the existing gap regarding the effectiveness of the different acoustic features for the recognition of emotions in speech. It presents a systematic analysis of the acoustic parameterisations that are used most for emotion identification (spectral, prosodic and voice quality characteristics), providing a complete description of their effectiveness for this task.

The purpose is to describe the effectiveness of isolated features as well as the behaviour of different sets of features. Therefore, individual parameterisations and their combinations have been analysed.

A special care was taken so that the obtained results can be comparable among different features. A common database and methodology was used all along the process in order to ensure this.

Furthermore, these features are supposed to work in real-life automatic emotion identification systems. Therefore all the parameterisation process had to be made completely automatic, without manual corrections. Some of the algorithms used during this process had to be modified, or new ones had to be developed, in order to make the process robust enough to work with natural emotions and spontaneous speech.

3. Methodology

3.1. Databases

The analysis of the features and the experiments were repeated using two publicly available databases. The first one, *Berlin EMO-DB* [13], is an acted emotional speech database, that contains recordings from 10 male and 10 female speakers. Each speaker repeated the same 10 sentences simulating seven different emotional states: anger, boredom, disgust, fear, happiness, neutral and sadness.

As it contains parallel corpora for each emotion, this database makes it easier to compare the different characteristics of the emotions. Therefore, it was used to make the first estimation of the discriminality of the parameterisations.

In order to validate the results obtained with acted speech, and to see whether the conclusions can be applied to natural emotions, a second analysis was carried out using the *FAU-Aibo* database [14]. This one contains spontaneous speech and natural emotions recorded from 21 boys and 30 girls while they played with the Sony Aibo pet robot. The database contains almost 18,000 recordings distributed in four speaking styles: anger, emphatic, neutral and positive.

3.2. Processing of the recordings

The recordings were processed in order to get the characteristic curves and labellings needed for the extraction of the features. The processing included detection of the vocal activity, estimation of the glottal source signal and of the intonation curve, voiced-unvoiced labelling, pitch-period marking and vowel detection.

All this processing was performed automatically without manual corrections. In order to obtain reliable results under these conditions, some new algorithms had to be developed, and others had to be modified. These algorithms included:

- A new *vocal activity detector* (VAD), based on the LTSE-VAD [15]. The algorithm was modified so that the result is independent from the SNR of the signal, which is an important factor when using spontaneous speech. The resulting algorithm is described in [16].
- A new F_0 *estimator* and *voiced-unvoiced labeller*. The intonation curve was computed with the *Cepstrum Dynamic Programming* (CDP) algorithm [17], which uses the cepstrum transform and dynamic programming in order to estimate the F_0 value and the voice-unvoiced labelling at once. The paper describing the algorithm also presents experiments comparing it to other well-known pitch estimators, concluding that CDP has the best robustness with low SNR signals.
- A new *vowel detector* based on a phoneme recogniser working with models of clustered phonemes [18]. The clustering provides a consistent and very robust set of models that achieves high detection accuracy.

3.3. Considered features

The presented analysis is focused on acoustic parameters that can be extracted directly from the speech signal without a recognition step: spectral envelope, prosodic and glottal flow features. The parameters are divided according to their temporal structure into segmental and supra-segmental features. The diagram in Fig. 1 presents a schematic view of the parameterisation process.

Segmental features describe the evolution of the parameter over time. LFPC values [5] were used as representative of the spectral envelope, whereas instantaneous F_0 and intensity values were selected as *prosody primitives*, i.e., instantaneous '*prosodic*' features.

Supra-segmental features collect long-term information, estimated over time intervals longer than a frame. In this work, this interval was defined as the time between two consecutive pauses, as detected by the VAD algorithm.

For the long-term characterisation of the spectrum, different statistics of the LFPC were estimated. Similarly, prosodic information was extracted in the form of long-term statistics of the prosodic primitives. Last, voice quality features were also considered, and they were again defined as long-term statistics of various glottal source signal parameters, which were estimated by inverse filtering of the speech.

The combinations of the different information types were studied at parameter level (applying early fusion techniques) as well as at classifier level (using late fusion techniques). This allowed to analyse the discrimination capacity of different temporal structures created with the same parameterisations.

3.4. Feature analysis methods

The emotion discriminality of each parameterisation was studied using various techniques, each one of them providing a different insight about the characteristics of the features. This analysis was carried out both for individual parameters as well as for combinations of features, in order to check whether these combinations were useful or not.

3.4.1. Inter-class and intra-class dispersion

The intra-class dispersion represents the width of the distribution for a certain feature or feature set and for a given emotion. Therefore, it is a measure of the variability of the parameterisation. The inter-class dispersion represents the separation among the emotions that a feature or set of features provides. The relation between both dispersions provides a measure of the overlapping of the class distributions, i.e., the confusion probability among the emotions. This relation can be estimated using the J_1 criterion:

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

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and S_W and S_B are the intra-class and inter-class dispersion matrices respectively. The more separated the features are for each emotion, the higher this value is. Therefore, J_1 values were computed for each feature family, in order to estimate their capability to discriminate emotions.

3.4.2. Unsupervised clustering

Unsupervised clustering aims to divide the feature vectors into clusters according to their distribution, so that vectors that are close to each other are assigned to the same cluster, and vectors that are far away are assigned to different ones. Given a set of parametrised emotional speech recordings, if the emotional classes are correctly separated with that parameterisation, the resulting clusters should correspond to each emotion. Generally speaking, the fewer clustering errors that occur, the better the discrimination is. So, the clustering error can be used as another measurement for the estimation of the emotion discrimination. For this purpose, a k-means clustering was performed for each feature family, and the results were compared.

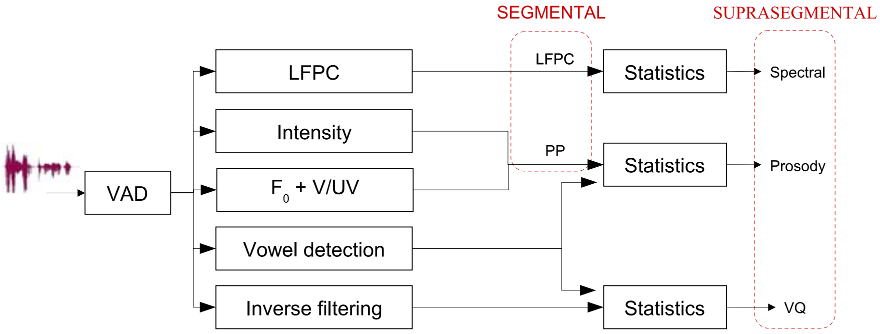

Figure 1: Schematic diagram of the parameterisation process.

3.4.3. *Feature selection*

A feature selection algorithm can help identifying the truly useful features, reducing the dimensionality of the parameterisation and making the classifier run faster and more accurately. Furthermore, detecting the discriminative features may provide a deeper understanding about the influence of the emotions in the acoustic characteristics of the voice.

The minimal-redundancy-maximal-relevance (mRMR) algorithm [19] was used to get a ranking of the features, from the most to the least significant one. mRMR was applied to all parameterisation families, as well as to their combinations, and the resulting rankings were carefully analysed in order to detect which features were most relevant individually and in combination with others.

3.4.4. *Experimental evaluation*

The final validation of the results obtained with the previous analysis was given by a series of experiments on automatic emotion identification using the various parameterisations and combinations of parameters that were considered. The experimental evaluation was carefully designed to be speaker independent, so that the results can be extrapolated to real-life conditions.

4. Results

4.1. Inter-class and intra-class dispersion

The J_1 values that were calculated show that spectral features provide larger separation among emotions than prosodic features. The emotion overlapping is highest when voice quality characteristics are used, which suggests that these kind of parameters are not suitable for automatic emotion identification.

As expected, the separation among the classes is smaller with real emotions and spontaneous speech than with acted speech. Nevertheless, the relation among feature sets is kept: spectral features separate emotions most and voice quality separates them least in both cases.

4.2. Unsupervised clustering

The results from the unsupervised clustering analysis are in accordance with J_1 values. Again, the number of clustering errors is larger with spontaneous speech than with acted speech. But in both cases the result using spectral features is better, with only a few recordings assigned to the wrong cluster, whereas the confusion is higher with prosodic features.

4.3. Feature selection

Results from the feature selection process are most interesting when the algorithm is applied to the combination of different feature sets. In these cases, the obtained ranking may show that the preferred features are of a certain nature, i.e., if spectral, prosodic or voice quality features are ranked in higher positions.

Regarding suprasegmental parameterisations, long-term spectral statistics are overall placed higher in the ranking than prosodic values. Voice quality features, instead, come out in the last positions. This effect is more evident with real emotions than with acted speech.

When the results of segmental parameterisations are analysed, it can be seen that short-term prosodic primitives are placed in quite good positions, although not in the best ones. In fact, when the ranking is performed over natural speech, prosodic primitives stay lower in the list, but still in a good place.

4.4. Experimental evaluation

The automatic emotion identification experiments confirm that features extracted from the spectral envelope of the speech are indeed more suitable for the task than parameters derived from the prosody or the voice quality. In fact, voice quality features have a really bad performance, and are almost of no use, even in combination with other kinds of parameters.

Prosodic characteristics, instead, may be useful if they are combined with spectral features. Nevertheless, this combination is relevant only in the case of acted speech. In the case of natural emotions, spectral features alone reach an accuracy similar to the one obtained with the combination.

5. Conclusions

The results from the analysis of the features reveal that, contrary to the most widely accepted theory, prosodic or voice quality features are not the most suitable ones for the automatic identification of emotions in speech. At least, not the kind of features that are typically used and that have been considered in this work. Spectral characteristics alone provide higher discrimination, and the combination of these spectral features with prosodic or voice quality ones does not improve the result. Although prosodic features may be useful at some extent when dealing with acted speech, they provide no accuracy improvement with natural emotions.

Most of the works that are focused on the analysis of features consider each parameter individually [20]. This way, they conclude that the variation from one emotion to another is larger for prosodic features than for spectral ones. But that is only applicable to individual parameters. The results from the inter-class and intra-class dispersion measures and the blind clustering suggest that, when the features are taken as a whole set, spectral characteristics provide more information about the emotion than prosodic ones.

The poor performance of prosodic and voice quality features is most probably due to the lack of robustness during their estimation from the speech signal. The low reliability of these features is more apparent with spontaneous speech. This effect is also shown in the described analysis, with prosodic and voice quality features performing better in acted speech than in spontaneous speech. The low reliability is far more noticeable for voice quality parameters, which are very difficult to obtain with enough robustness unless a manual supervision is applied.

It is not that features extracted from prosody or voice quality are useless. Several papers show that humans are able to identify emotions in prosodic copy-synthesis experiments [21], confirming that prosody does carry a great amount of emotional information, at least for some emotions. But the traditional prosodic representations may not be well suited to capture this information. On the one hand, long-term statistics estimated over the whole sentence lose the information of specific characteristic prosodic events. On the other hand, short-term prosodic primitives do not capture the prosodic structure correctly, which is suprasegmental by definition. The results suggest that a new more elaborate representation is needed to effectively extract the emotional information contained in the prosody, and that new and more robust algorithms are needed in order to capture this information with enough robustness.

Finally, the behaviour of the different sets of parameters for the discrimination of emotions has been similar both using acted and natural speech, with the expected loss of performance due to the more challenging problem that spontaneous speech poses. Therefore, acted speech databases can be used to perform the feature selection process and tune the automatic emotion identification system even if it has to deal with real emotions.

A paper describing part of this dissertation, and with the title "*Feature analysis and evaluation for automatic emotion identification in speech*" has been recently published in the IEEE Transactions on Multimedia, vol. 12, pp.490-501.

6. Acknowledgements

The author would like to thank the University of the Basque Country (UPV/EHU) for providing the grant and funding necessary to complete this work.

7. References

[1] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotional recognition," *Signal Processing*, vol. 88, pp. 2956–2970, Dec. 2008.

[2] S. Casale, A. Russo, and S. Serano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Communication*, vol. 49, no. 10, pp. 801–810, Aug. 2007.

[3] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, 2005.

[4] I. Luengo, E. Navas, I. Hernáez, and J. Sanchez, "Automatic emotion recognition using prosodic parameters," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 493–496.

[5] 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.

[6] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional speech," in *Speech evaluation in Psychiatry*, J. K. Darby, Ed. New York, USA: Grune and Stratton, 1981, pp. 189–220.

[7] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Pathology*, vol. 70, no. 3, pp. 614–636, 1996.

[8] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," in *IEEE Workshop on Multimedia Signal Processing*, Crete, Oct. 2007, pp. 48–51.

[9] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," *Lecture Notes on Computer Science*, vol. 4738, pp. 139–147, 2007.

[10] E. Navas, I. Hernáez, I. Luengo, I. Sainz, I. Saratxaga, and J. Sanchez, "Meaningful parameters in emotion characterisation," *Lecture Notes on Artificial Intelligence*, vol. 4775, pp. 74–84, 2007.

[11] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man. automatic classification of emotions and inter-labeller consistency," in *ICASSP*, Philadelphia, USA, Mar. 2005, pp. 317–320.

[12] I. Luengo, E. Navas, and I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge," in *Interspeech*, Brighton, UK, Sep. 2009, pp. 332–335.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, Lisbon. Portugal, Sep. 2005, pp. 1517–1520.

[14] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Information Society - Language TechnologiesConference (IS-LTC)*, Ljubljana (Slovenia), Oct. 2006, pp. 240–245.

[15] 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.

[16] I. Luengo, E. Navas, and I. Hernáez, "Modified LTSE VAD algorithm for applications requiring reduced silence frame misclassification," in *Language Resources and Evaluation Conference (LREC)*, 2010, p. (To appear).

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

[18] I. Luengo, E. Navas, J. Sánchez, and I. Hernáez, "Detección de vocales mediante modelado de clusters de fonemas," *Procesado del Lenguaje Natural*, vol. 43, pp. 121–128, Sep. 2009.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[20] E. Navas, I. Hernáez, A. Castelruiz, J. Sánchez, and I. Luengo, "Acoustic analysis of emotional speech in standard Basque for emotion recognition." *Lecture Notes on Computer Science*, vol. 3287, pp. 386–393, Oct. 2004.

[21] E. Navas, I. Hernáez, and I. Luengo, "An objective and subjective study of the role of semantics in building corpora for emotional TTS," *IEEE transactions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1117–27, Jul. 2006.