Speech Emotion Recognition Using Hidden Markov Models

Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño

Research Center TALP, Universitat Politècnica de Catalunya. SPAIN.
{albino,asuncion,antonio,canton}@gps.tsc.ups.es
http://gps-tsc.upc.es/veu/

Abstract

This paper introduces a first approach to emotion recognition using RAMSES, the UPC's speech recognition system. The approach is based on standard speech recognition technology using hidden semi-continuous Markov models. Both the selection of low level features and the design of the recognition system are addressed. Results are given on speaker dependent emotion recognition using the Spanish corpus of INTERFACE Emotional Speech Synthesis Database. The accuracy recognising seven different emotions—the six ones defined in MPEG-4 plus neutral style—exceeds 80% using the best combination of low level features and HMM structure. This result is very similar to that obtained with the same database in subjective evaluation by human judges.

1. Introduction

Dealing with the speaker's emotion is one of the latest challenges in speech technologies. Three different aspects can be easily identified: speech recognition in the presence of emotional speech, synthesis of emotional speech, and emotion recognition. In this last case, the objective is to determine the emotional state of the speaker out of the speech samples. Possible applications include from help to psychiatric diagnosis to intelligent toys, and is a subject of recent but rapidly growing interest [1].

This paper describes the TALP researchers first approach to emotion recognition. The work is inserted in the scope of the INTERFACE project [2]. The objective of this European Commission sponsored project is "to define new models and implement advanced tools for audio-video analysis, synthesis and representation in order to provide essential technologies for the implementation of large-scale virtual and augmented environments. The work is oriented to make man-machine interaction as natural as possible, based on everyday human communication by speech, facial expressions and body gestures."

In the field of emotion recognition out of speech, the main goal of the INTERFACE project will be the construction of a real-time multi-lingual speaker independent emotion recogniser. For this purpose, large speech databases with recordings from many speakers and languages are needed. As these resources are not available yet, a reduced problem will be addressed first: emotion recognition in multi-speaker language dependent conditions. Namely, this paper deals with the recognition of emotion for two Spanish speakers using standard hidden Markov models technology.

2. Recognising Emotion in Speech

Several studies show a high correlation between some statistical measures of speech and the emotional state of the speaker [3, 4, 5, 6]. Among these measures, the most popular are related to pitch, energy, articulation and spectral shape. For instance, sadness have been associated to low standard deviation of pitch and slow speaking rates, while anger usually implies higher values of pitch deviation and rate. Most of the efforts done so far in emotion recognition are based on determining which are these sources of information and how can we deal with them. The first question: where the information is, is usually answered in terms of statistical measures of the utterance. An extensive enumeration of the so far proposed measures can be found in Tables 6-10 of a recent work from Cowie *et al.* [1]. Most of the measures are means, medians, standard deviations, or percentiles, estimated over the whole utterance. With respect to how can we deal with these kinds of measure to get the information about the emotional state, their own nature aim at the use of knowledge driven algorithms and/or discriminant analysis. Accuracies in the order of 50%—close to the performance of human judgers— are reported in the above mentioned work showing the utility of the approach.

Nevertheless, taking global statistics also present several drawbacks. In first place, it must be considered the fact that they ignore the temporal structure of speech, being sensible to properties that could otherwise be linguistically determined. The phonetic contents of the utterance and its structure play many times a role as important or more as emotion does. For instance, interrogative sentences usually imply a wider pitch contour than affirmative ones, thus their pitch standard deviation will usually be higher. But this has nothing to do with the emotional style, only with the sentence nature. Another limitation of using global statistics is the fact that the processing can only be done once the whole utterance has been pronounced. This fact limits the capability of building real time recognisers and is a main drawback when the emotion varies along the utterance.

A different approach to global statistics is taking into consideration that this kind of modeling is just a reflection of the short time behaviour. For instance, instead of using means and standard deviations of short time raw features as energy or pitch, we can deal directly with their probability distribution function (**pdf**). Yet, the pdf of a population carries the same information as those *cooked* measures, while providing a much more precise description of the population. If we consider the case of modeling the pdf with a mixture of Gaussian distribution, the problem is equivalent to that of using hidden Markov models (**HMM's**) of just one state. Hidden Markov models have a long tradition in speech recognition. The underlying idea is that the statistics of voice are not stationary. Instead of that, voice is modeled as a concatenation of *states*, each of which models different sounds or sound combinations, and has its own statistical properties. There are two main advantages of HMM's in front of global statistics for emotion recognition: first, the structure of HMM's may be useful to catch the temporal behaviour of speech; sec-

ond, HMM technology has been long time studied for speech recognition purposes, being available well established procedures for optimising the recognition framework: Baum-Welch algorithm, discriminative training, etc.

3. Low Level Features for Emotion Recognition Using HMM's

The first problem that arises when trying to build a HMM based recognition framework is the selection of the features to be used. In this case, it is not enough that the feature carries information about the emotional state, but it must fit the HMM structure as well. The main consequence of this limitation is that the features used must model the short time behaviour of voice, invaliding the use of global statistics estimated from the whole utterance.

In our work we have decided to use those raw features that could lead to statistical measures more similar to those proposed in literature for emotion recognition [1]. We first considered the following possible features: short time pitch and energy, the contours of pitch and energy, the spectral shape, and duration and silence related measures. Of these measures we have only tried four: the instant values and contours of pitch and energy. These features are easy to estimate in real time frameworks, while being known to carry a large amount of information about the emotional state. Their means, standard deviations, etc. are measures proposed in almost every work related to emotion recognition. Besides, if careful chosen, pitch and energy features can be made quite robust to channel distortion, speaker, sex and even language.

Spectral measures were discarded in our first approach to emotion recognition because they need complex frameworks to be characterised. This is so because the spectrum depend heavily on the phonetic content of the sentence. Pitch and energy also do, but we can expect them to depend only on broad classes of sounds, rather than on phonemes. Another reason for discarding spectral measures is our believe that their phonetic dependency would be a main drawback for building language independent emotion recognisers.

Although many often named, we also discarded direct use of temporal and silence related measures because they need a previous recognition step in order to get a phone/silence segmentation/recognition, increasing the complexity of the overall system. Yet, the HMM structure along with a good choice of the pitch and energy features can provide a quite good representation of this kind of measures. Articulatory rate, and frequency and duration of silences, for instance, will have direct implications in the pdf of pitch and energy, and their derivatives.

Absolute values and long term evolution of some parameters are avoided due to their dependency on factors that have nothing to do with the speaker's emotional state. For instance, the absolute value of energy reflects not only the intentional level, but the sex and age of the speaker and the gain of the recording chain as well. On the other hand, whether a sentence is affirmative or interrogative, or its length, will probably play a determinant role on the whole sentence contour of pitch. For both energy and pitch, we consider two kinds of temporal scope: instantaneous values and syllabic contour. In the first case, raw low level analysis is performed on samples of 25ms taken every 10ms. In the second, these same features are processed in order to capture their mean behaviour in segments of 100ms—that roughly correspond to the length of two phones—.

3.1. Energy features

In order to model the instantaneous values of energy without relying on the absolute value of energy we use the first and second derivatives of the logarithm of the mean energy in the frame. The acoustical meaning of these measures is related to the sharpness of the energy level, reflecting both the articulation speed and the dynamic range. Besides, effects such as tremor—small frequent variations in voice intensity—are also easily characterised by the instantaneous energy levels.

Syllabic contour of energy is modeled by the first and second derivatives of the logarithm of the 8Hz low pass filtered energy in the frame. In this case, the relative intensity of consecutive sounds will be represented.

3.2. Pitch features

Pitch features present a similar behaviour as energy ones. In this case we are neither interested in the global pitch —which is heavily influenced by the speaker's nature—, nor its global evolution along the utterance —which will depend on the sentence structure—. Besides, we can expect the syllabic contour of pitch and its instantaneous levels to provide profitable information about the emotion.

In order to characterise instantaneous pitch, a simple auto-correlation analysis is performed at every frame. The maximum of the long term auto-correlation is determined and used to form five different parameters: the value of the maximum of the long term auto-correlation, along with its first and second derivatives; and the first and second derivatives of the logarithm of the pitch lag.

The raw auto-correlation maximum is a measure of the *harmony* of voice. High values of this maximum imply a high periodicity in the speech waveform, while low values imply low or none periodicity. This feature allows us to discern between harsher styles—such as anger or disgust—and other more musical styles—as joy or surprise—. Besides, the first derivative of the logarithm of pitch lag represent the relative variation of pitch between frames. In order to model the pitch lag, we use the position of the maximum of the long term auto-correlation, without further processing. Thus, it will present abundant errors, particularly pitch lag doubling and halving. These artifacts will be characterised in the first derivative of the logarithm of the pitch lag as fixed constants of $\pm \log 2$. Thus, the first derivative of the logarithm of the pitch lag will help to detect those styles for which these effects are more frequent The derivatives of the logarithm of the pitch lag are also expected to help in the detection of jitter, the presence of fast fluctuations in the very short time values of pitch.

The syllabic contour of pitch is characterised with a twice filtered version of the pitch lag estimated for instantaneous pitch: first the pitch lag is median filtered in order to remove artifacts from the estimation; second an 8Hz low pass filter is applied to capture the syllabic contour. In this case a much more precise estimation of pitch is used in order to capture the actual evolution of its values and not the frequency of errors in the estimation. The first and second derivatives of the smoothed pitch lag evolution obtained this way are then evaluated in order to represent the pitch evolution in segments of a few phones.

4. Experimental Framework

4.1. INTERFACE Emotional Speech Synthesis Database

The INTERFACE Emotional Speech Synthesis Database, IESSDB [7], was recorded in four different languages, French, English, Slovenian and Spanish. Each language contains utterances from two professional actors (one male and one female) simulating each of the six MPEG-4 defined emotional styles—anger, disgust, fear, joy, sadness and surprise—plus a supplementary neutral one. The design of the database is specially oriented to speech synthesis purposes, but it can also provide a first approximation to emotional speech analysis and emotion recognition. The recordings were all done in silent rooms and with high quality microphones—an AKG 320, in the Spanish database—. Finally, the utterances were sampled with 16 bits, first at high rates—32KHz, in the Spanish case—and then down sampled to 16KHz.

Six different kinds of sentence were recorded: affirmative, exclamatory and interrogative sentences, paragraphs of around five sentences, and isolated words and digits. Two different sessions were recorded with each speaker, with more than two weeks elapsed between them. In the case of the Spanish database, the affirmative sentences and the paragraphs were also uttered in just one of the sessions in neutral style but at high and low rates, and loud and soft intensities.

4.1.1. Subjective Evaluation of IESSDB

An informal subjective evaluation of IESSDB was carried out with 16 non professional listeners [8]. A total of 56 utterances were played simultaneously to all the listeners. The utterances, seven per emotion, include equal proportion of both speakers and sessions. Half of them are digits and isolated words, and the other half are sentences and paragraphs. Each listener had to choose between the seven emotional styles considered, valuing the intensity of the perception in an one to five scale. If the listener was not convinced of his choice, he could also mark a second one.

The results of the evaluation were quite satisfactory: more than an 80% of the first choices were correct, and this figure almost reaches 90% if second choices—which were very scarce—are admitted. Moreover, in all 56 utterances more than a 50% of the listeners chose the correct emotion. Besides, all errors were committed for digit and isolated word utterances. All the sentences and paragraphs were correctly recognised in first choice by all the listeners. No notorious difference was observed in the results neither between one speaker and another nor between sessions.

Table 1 shows the confusion matrix of the experiment. Errors mainly affect two different sets of emotions: first, fear, disgust and sadness; second, surprise, joy and anger. Most of the errors committed involved emotions of either of the two sets, which is usually confused with another emotion of the same set.

Table 1 shows the confusion matrix of the experiment. There are two main kinds of error: first, confusions between emotions inside either the fear-disgust-sadness set, or the surprise-joy-anger one; second, recognising neutral style—in a way detecting no emotion at all—, instead of any of the other 6 ones.

These results serve to validate the characterisation of the seven emotions by the two speakers. From a voice synthesis point of view, the goal has been reached. For speech recognition purposes, the database also seems to be adequate: it is clear in advance that there is information about the emotion in each of the utterances. Particularly in the material that will be used in the experimentation presented in this paper: the sentence utterances. Obviously, being professional actors, emotion marks are simulated and may be overdone, so recognition results in real conditions may be very different from those obtained with these two actors. Nevertheless, it seems a plausible starting point for emotion recognition research.

	S	**J**	**A**	**F**	**D**	**T**	**N**
Surprise	89	20	7	0	6	2	4
Joy	0	115	7	0	2	2	2
Anger	2	14	85	2	5	5	15
Fear	4	1	1	103	5	13	1
Disgust	2	1	2	5	106	3	9
Sadness	1	3	1	16	3	101	3
Neutral	0	2	2	1	4	1	118
Total	**98**	**156**	**105**	**127**	**131**	**127**	**152**

Table 1: Confusion matrix of the emotion subjective evaluation of IESSDB. Columns represent the emotion elected in first choice for utterances belonging to the emotion of each row, where A stands for anger, S for surprise, J for joy, F for Fear, D for disgust, T for sadness and N for neutral. The number of utterances per emotion is, in all cases, 128, with a grand total of 896.

4.1.2. IESSDB subset for multi-speaker emotion recognition evaluation.

In order to carry out the experimentation, we have only used a part of the Spanish corpus of IESSDB: the 100 affirmative and phonetically balanced sentences uttered in the six MPEG-4 styles plus neutral. Both sessions of the two speakers are used. This leads to a total of $2 \times 2 \times 7 \times 100 = 2800$ utterances (actually 2782) which were divided in two non overlapping sets: one with 2227 utterances for training purposes, and another with 555 utterances for test. Both sets are designed in such a way that the contents of each speaker, session, emotion and sentence is maximally balanced across the training and testing parts of the corpus.

4.2. Using RAMSES in Emotion Recognition

Our proposal consists in modeling short time low level features with semi-continuous HMM's. In a first approximation, we just use one HMM per emotion. As low level features we study the performance of the four different sets of parameters described in Section 3. For each set of parameters different choices of the number of states of the HMM's are tried. An experiment is also carried out with all the low level features together and different number of states. In this experiment all 11 low level parameters used to form the above studied 4 classes are mixed together assuming independence in the information provided by each of them.

11 different low level features are estimated at 100 frames by second. Each of the low level features is quantified with a Gaussian codebook of 64 codewords and used to train seven different HMM's, one per emotion. In the recognition phase the maximum likelihood HMM is selected for each utterance. To compare results with global statistics frameworks, single state HMM's are first tried. Single state HMM's are a way to represent the probability distribution function of each feature by

Feature	Number of States 1	8	16	32	64
InstEner	36.9	38.8	43.2	47.4	48.8
SyllEner	34.4	37.3	40.7	42.9	42.3
InstPtch	57.9	65.8	68.5	72.6	75.5
SyllPtch	45.8	48.3	48.3	47.7	47.9
EnerPtch	60.4	73.9	81.4	81.4	82.5

Table 2: Accuracy in the emotion recognition of the Spanish Corpus of IESSDB. Results are shown for different sets of features and number of states of the HMM's. InstEner stands for instantaneous energy, SyllEner for syllabic contour of energy, InstPtch for instantaneous pitch and SyllPtch for syllabic contour of pitch. Finally, EnerPtch stands for the overall set.

means of a mixture of Gaussian distributions. In our opinion, this is a robust alternative to some of the global statistic measures used so far in emotion recognition—means, variances and medians, for instance—. Several numbers of states—8, 16, 32 and 64—are also tried for each low level feature, showing the effects of the increase of temporal complexity.

4.3. Emotion Recognition Results

Table 2 shows the percentage of correctly identified emotions for the 555 utterances of the test set using RAMSES as emotion recognition engine. Results are given for five different combinations of low level features: the four classes discussed in Section 3, and the overall set of them all.

The first remarkable conclusion about these results is the fact that in all cases the low level feature seems to be useful for emotion recognition. Using single states HMM's—somehow equivalent to taking global statistics—gives recognition results well above the double of chance level—14,3%, for seven emotions—. Instantaneous features provide better performance than syllabic ones for both energy and pitch, and pitch features work also better than energy ones. As a result, the best partial combination is instantaneous pitch. Being the worst the syllabic contour of energy. Nevertheless, the best combination at all is the complete 11 features set, which reports the highest accuracy independently of the number of states. All the sets of features show a noticeable improvement when bigger models are used, although there is certain saturation for HMM's of more than 32 states.

Table 3 shows the confusion matrix for the best combination of low level features and HMM's number of states: all 11 features mixed together in 64 states HMM's. Besides its similarity with the subjective results (depicted in Table 1)—in this case, most errors involve the surprise-joy-anger-fear set, and the recognition of neutral instead of any of the other 6 styles—, it is remarkable the homogeneity in the results: all emotions are recognised with an accuracy higher than 70%.

5. Discussion

In this paper, an HMM based approach to emotion recognition has been presented. Results—an accuracy higher than 80% in the speaker dependent recognition of seven emotional styles—confirm both the usefulness of the approach and the convenience of the low level features used. Given the reduced scope of the scenario considered it may be arguable if the results can be generalised to other speakers and/or languages, yet we believe that the results achieved are encouraging: at least, they show the usefulness of the approach for multi-speaker emotion recognition Besides, we believe that this is a good baseline for more ambitious tasks.

	S	J	A	F	D	T	N	Tot
Surprise	68	4	1	5	1	0	0	**79**
Joy	11	56	9	3	1	0	0	**80**
Anger	4	9	59	0	4	0	3	**79**
Fear	8	3	0	60	4	1	4	**80**
Disgust	0	2	2	1	72	0	2	**79**
Sadness	0	0	0	0	0	64	15	**79**
Neutral	0	0	0	0	0	0	79	**79**
Total	**91**	**74**	**71**	**69**	**82**	**65**	**103**	**555**

Table 3: Confusion matrix of the emotion recognition using all four kinds of low level features together and HMM's of 64 states. Columns and rows have the same meaning as for Table 1.

In the experimentation a new database was used: INTERFACE emotional speech database. Results—either of subjective tests, or using the framework proposed in this paper—show the utility of this database for its main purpose, emotional speech synthesis, and as a starting point for emotion recognition research.

Future work includes, first of all, working in speaker—and probably language—independent conditions. Besides, we plan to carry out a larger research on low level feature extraction, alternative acoustic modeling techniques, and multi-modal emotion recognition.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] INTERFACE Project, "Multimodal analysis/synthesis system for human interaction to virtual and augmented environments," EC IST-1999-No 10036, coor. F. Lavagetto, 2000–2002, http://www.ist-interface.org.

[3] B. Heuft, T. Portele, and M. Rauth, "Emotions in time domain synthesis," in *Proc. of ICSLP*, Philadelphia, Oct. 1996, pp. 1974–1977.

[4] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. of ICSLP*, Philadelphia, Dec. 1998, pp. 1989–1992.

[5] N. Amir and S. Ron, "Towards an automatic classification of emotion in speech," in *Proc. of ICSLP*, Sydney, Dec. 1998, pp. 555–558.

[6] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "Acoustic nature and perceptual testing of corpora of emotional speech," in *Proc. of ICSLP*, Sydney, Dec. 1998, pp. 225–228.

[7] C. Tchong, J. Toen, Z. Kacic, A. Moreno, and A. Nogueiras, "Emotional speech synthesis database recordings," Tech. Rep. IST–1999–No 10036–D2, INTERFACE Project, July 2000.

[8] I. Hernández, "Diseño de un corpus para síntesis de voz en castellano," Graduation thesis, dir. A. Moreno, Universitat Politècnica de Catalunya, Barcelona, 2000.