Extracting emotions from speech signal: State of the art - Seminar Paper[*]

Sierro Hervé
University of Fribourg
1700 Fribourg
Switzerland
herve.sierro@unifr.ch

ABSTRACT

Nowadays we can observe more and more sophisticated procedures to extract emotions from a speech signal and to classify them. Emotional speech recognition can be separated as following: select/build a speech data collection, extract features from the speech signal and select/build a classifier algorithm. This article will give an overview of the state-of-the-art in the field, will bring out what are the most important features to be taken into account for each steps of emotions recognition and will give an account of the lacunas in this field. The most important finding is the lack of data for research on spontaneous/real-life speech, both in terms of data collections and features.

General Terms

Speech emotions recognition

Keywords

Emotion; Speech signal; Data speech collection; Feature types; Evaluation

1. INTRODUCTION

The interest for emotional speech recognition has grown considerably during the past ten years (>100 papers per year since 2004)[1]. "Enabling future non-human interaction partners - be this computers, robots, or something else - at least to come closer to the normal human abilities to recognise 'emotions' and to respond in an emotionally intelligent way" [1], or be able to understand how our own emotions are working, are two of the many goals of this research area.

This article is structured as follows : we first deal with data speech collections, analysing existing databases, their features and their lacks including discussions on prompted/non-prompted emotions and corpus of emotions. From there we go to the features section in where we describe how features are classify, give an overview of main features and discuss on selection methods. Finally we briefly introduce classification algorithm dealing with static and dynamic classifiers.

2. DATABASES

When we talk about emotional speech recognition, the first thing we must deal with is the selection or the creation of a data speech collection.

An overview of 64 emotional speech data collections (from 1986 to 2006) is presented in [2]. For each data collections, the following characteristics are given: speech language, number and profession of subjects, other physiological signals possibly recorded simultaneously with speech, data collection purpose (emotional speech recognition, expressive synthesis), emotional states recorded, and the kind of emotions (natural, simulated, elicited).

The important informations are certainly how emotions are produced and the corpus of emotions recorded. This will be detailed below.

2.1 Modelling emotions

The most represented emotions by the overview of 64 emotional speech data collections in [2] are anger, fear, sadness, joy, or surprise and are known as the *big n* emotions category. Indeed, emotions can be modelled as categories which can be separate from *main categories* - such as positive, neutral, negative or the *big n* emotions, e.g., anger, fear, sadness, joy, etc. - to *sub-categories* modelling shades of the mains categories, e.g., surprised, irritated, bored, etc.

Another way to model emotions is the dimension's aspects where in it is foremost the question of how many dimensions will be taken into account. Human emotional states can be represented in three dimensional spaces: "Arousal is the individual's global feeling of dynamism or lethargy. It subsumes mental activity as well as physical, preparedness to act as well as overt activity. The Power dimension subsumes two related concepts, power and control. However, people's sense of their own power is the central issue that emotion is about, and that is relative to what they are facing. Valence is an individual's overall sense of "weal or woe": Does it appear that on balance, the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state?" [10]

Experiments using only prosodic features suffer from a clear weakness, trying to differentiate emotions, which are placed very close in the arousal dimension, but in the contrary being quite separated in the valence dimension are reported in

[*] Seminar emotion recognition : http://diuf.unifr.ch/main/diva/teaching/seminars/emotion-recognition

[1] based on a search for 'emotion and speech and recognition' in Scopus

[3]. So, considering the three dimensional spaces turns out to be useful to recognize emotions.

2.2 Type of the speech

When a database is created, we first have to deal with the problem of *prompted* or *non-prompted* emotions [1].

By *prompted* emotions, we mean that speakers have to produce given specific emotions while often reproduce segmentally identical utterances. Such production of emotions are usually performed by professional actors. However, when the adequate number of professionals is not available, non-professionals while imitating a professional can be accepted and in this case we will talk about *elicited speech.* A research highlights the fact that acted speech from professionals is the most reliable because professionals are trained to color a speech by emotions and such emotions have a great amplitude or strength [2]. But the disadvantage of *prompted* emotions is the fact that deliberately acting emotions, is not like producing a spontaneous emotion. Indeed, we can never be sure, even if we work with actors, that the emotion produced will be the same than placing the author in the context of the real life. In order to avoid this situation, a context can be sometimes given by looking at video clips or thinking about specific situations. Despite these solutions, we can never be sure that the result will be natural and remain monolithic and not interactive.

By *non-prompted* emotions, we mean that subjects do not reproduce specific emotions but emotions are confined to natural scenarios. Indeed, in real life, dialogs and interactions between speakers can contain a lot of emotions. However, assigning one emotion to one situation seems difficult because firstly, emotions are subjective and secondly, in natural situations emotions can be contrasted. For this reason, in *non-prompted* speech, we speak more about emotion-related states. Recording such scenarios can be done in two ways: it may either consist of an interaction between humans (*human-human*) or an interaction between a human and a machine (*human-machine*) where the machine can be a computer or a robot that can eventually be controlled by a human operator.

Today we can note that most emotional speech databases include prompted emotions. Indeed, if we analyse the emotional speech data collection in [2], we count 37 prompted, 18 natural and 9 elicited. These numbers are not surprising because simulated emotions were way easier to obtain than natural ones, since emotions are harder to label in real life. As mentioned above, *prompted* emotions are not ideal to represent real life. Additionally, acted emotions are more easily recognized than realistic emotions [4]. Accordingly, ideal way to record emotion is to record speakers who do not know that they are being recorded but this solutions generates private rights issues. Finally, the solution is to design scenarios that are as close as possible to real life.

3. FEATURES

The previous section gives an overview of the necessary prerequisites for building speech data collections. In this section, we address the second important step in emotions recognition that is the extraction of emotional relevant features from speech.

There is no unique ways to classify features but preferable way towards the following taxonomy: acoustic and linguistic features are usually considered separately due to the extreme difference concerning their extraction methods. Another distinction is made upon of the database used. Indeed, for spontaneous/real-life speech, linguistic features can considerably gain in importance but for acted speech, these features loose their value, since utterances are identical for all speakers.

Few years ago, just a small set of features were used. Today, the large number of acoustic feature (low-level-descriptor LLD) and functionals (rich statistical description of LLD) has newly supported the extraction of very large feature vectors, up to many thousands of features [1]. In the following, explanations of acoustic/linguistic features and functionals are detailed.

Figure 1 provides a taxonomy of features commonly used for acoustic and linguistic emotion recognition.

3.1 Acoustic features

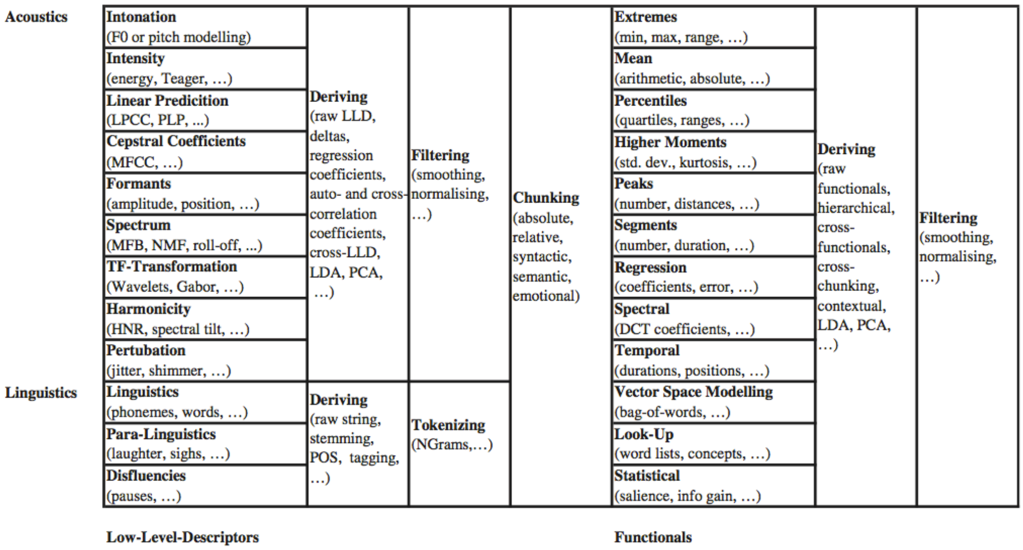
Acoustic feature are not meaningful for emotions, but rather their behaviour over time. Consequently, commonly statistics such as minimum, maximum or mean from time series of these measures are calculated. In this way, the time series of values have to be fragmented into parts from which to compute the statistics. A study concerning a future on-line emotion recognition system demonstrates that a considerable decrease in recognition accuracy can be observed when segment length gets shorter with acted emotions [4]. In fact, the *whole utterance* reaches the best ratio, followed by *word in context* (word with its leading and subsequent word), a fix segment length of *500ms* and finally by *word.* For non-acted emotions, the longer unit again comes off better, but the difference is not as striking due to the fact that phrase and word contours are less distinct in spontaneous speech.

Acoustic features are for example : pitch, intensity, duration, voice quality, etc. In the literature, we often found the terms *prosodic features* and *voice quality* which are part of acoustic features. The mainly used acoustic features are derived from speech processing, whereas the prosody is characterized by large statistics measures of pitch, energy and duration [11]. The idea behind using acoustic features for emotion recognition is based on the fact that humans use voluntary or involuntary acoustic variation to mark the importance of particular items in their speech [5].

Duration features model temporal aspects. Different types of normalisation can be applied : because pitch, energy or the duration of voiced and unvoiced segments are measured in seconds, they can represent duration features. In other hand, it can exclusively represent "the parameter 'duration' of higher phonological units like phonemes, syllables, words, pauses or utterances" [1].

Intensity features "model the energy of a sound as perceived by the human ear, based on the amplitude in different intervals" [1] and refers to the strength of a sound wave. Energy features depend upon both intensity and frequency and can model intervals or characterising points.

The **pitch** is usually taken to be the fundamental frequency *F0* that is defined as the lowest frequency of a periodic waveform which it is measured in Hz. The pitch is generally computed by auto-correlation methods from the speech signal and carries information about emotion because it depends on the tension of the vocal folds and the sub-glottal air pressure. Pitch extraction is error-prone it-self but the experimental results indicate that the influence is rather small, at least for the current state-of-the-art in modelling pitch



	Low-Level-Descriptors				Functionals		
Acoustics	**Intonation** (F0 or pitch modelling)	**Deriving** (raw LLD, deltas, regression coefficients, auto- and cross-correlation coefficients, cross-LLD, LDA, PCA, …)	**Filtering** (smoothing, normalising, …)	**Chunking** (absolute, relative, syntactic, semantic, emotional)	**Extremes** (min, max, range, …)	**Deriving** (raw functionals, hierarchical, cross-functionals, cross-chunking, contextual, LDA, PCA, …)	**Filtering** (smoothing, normalising, …)
	Intensity (energy, Teager, …)				**Mean** (arithmetic, absolute, …)		
	Linear Predicition (LPCC, PLP, ...)				**Percentiles** (quartiles, ranges, …)		
	Cepstral Coefficients (MFCC, …)				**Higher Moments** (std. dev., kurtosis, …)		
	Formants (amplitude, position, …)				**Peaks** (number, distances, …)		
	Spectrum (MFB, NMF, roll-off, ...)				**Segments** (number, duration, …)		
	TF-Transformation (Wavelets, Gabor, …)				**Regression** (coefficients, error, …)		
	Harmonicity (HNR, spectral tilt, …)				**Spectral** (DCT coefficients, …)		
	Pertubation (jitter, shimmer, …)				**Temporal** (durations, positions, …)		
Linguistics	**Linguistics** (phonemes, words, …)	**Deriving** (raw string, stemming, POS, tagging, …)	**Tokenizing** (NGrams,…)		**Vector Space Modelling** (bag-of-words, …)		
	Para-Linguistics (laughter, sighs, …)				**Look-Up** (word lists, concepts, …)		
	Disfluencies (pauses, …)				**Statistical** (salience, info gain, …)		

Figure 1: Taxonomy of features commonly used for acoustic and linguistic emotion recognition.

features and for this type of data [2].

Voice quality models many various measures of voice quality and therefore is a complicated issue in itself. Measures of the quality of the speech signal are noise-to-Harmonic Ratio, shimmer, jitter, and further micro-prosodic events. "Although they partially depend on other LLDs such as pitch (jitter) and energy (shimmer), they reflect peculiar voice quality properties such as breathiness or harshness." [1]

The **spectrum** of speech reveals information on the formants, that are one of the quantitative characteristics of the vocal track. Formants are characterized by their frequency and their bandwidth. With formants, it is possible to detect if the speech was articulated or slackened because formant bandwidth during slackened articulated speech is gradual, whereas the formant bandwidth during improved articulated speech is narrow with steep flanks [2].

The **cepstrum** is the result of taking the Fourier transform of the logarithm of the spectrum of a signal and it emphasises changes or periodicity in the spectrum, "while being relatively robust against noise" [1].

Wavelets are waves like oscillation with amplitude that starts out at zero, increases, and then decreases back to zero. "They give a short-term multi-resolution analysis of time, energy and frequencies in a speech signal." [1]

3.2 Linguistic features

The words we choose or the grammatical alternations that we use, also play a role in the reflection of our emotional state. A number of techniques exist for this analysis but the two predominant methods are N-Grams and Bag-of-Words. The first method is based on a probabilistic language model for predicting the next item in a given sequence, whereas the second is a well-known numerical representation form of texts in automatic document categorisation.

Before applying this technique, it is useful to reduce the complexity of the speech. For this purpose, approaches such as elimination of irrelevant words (expert-based list of words) or stopping words that do not exceed a general minimum frequency of occurrence, are used.

Tokenisation can be obtained by mapping the text into word classes. A first popular choice is *lexemes*, called Stemming. It consists of clustering words to their stem and reduces the number of entries in the vocabulary. A second choice is part-of-speech where classes such as nouns, verbs, adjectives or more detailed sub-classes are modelled. Finally, semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms. Non-linguistic vocalisations (sighs, laughs, cries, etc) can also easily be integrated into the vocabulary.

3.3 Functionals

Functionals is the fact of applying after LLD (Low-level-descriptor) extraction, a number of operators and functionals "to obtain feature vectors of equal size out of each base contour" [1]. Functionals provide a sort of normalisation over time: with the usage of those, one feature vector per word with a constant number of elements can be obtain, ready to be modelled by a static classifier.

As for linguistic features, LLD can be filtered or transformed before functionals are applied: first or second derivatives are oftentimes calculated and finish as additional LLDs. Functionals features are for example the four first moments (mean, standard deviation, skewness and curtosis), extremes values (min, max, range), higher moments, peak (number, distance) or segments (number, duration).

However, functionals are not always necessary. Dynamic classifiers, such as HMMs (Hidden Markov Models) process correctly the LLDs to classify them and provide implicit time normalisation.

3.4 Features selection

The aim of the features selection is to select a subset of features to describe a phenomenon from a larger set that may contain irrelevant or redundant features. Improving classifier performance and accuracy are usually the motivating factors behind features selection [6].

Widely used general approaches are *wrapper based selection* methods which employ a target classifier's accuracy as optimisation criterion in a 'closed loop' fashion [1] where a feature that has a poor performance will not be conserved. Probably the most common chosen procedure is the sequential forward search, a hill climbing selection starting with an empty set and sequentially adding and keeping features that allow an improvement of performances. Second general approaches are *filters* methods which totally ignore the effects of the selected feature subset on the performance [7].

Finally, the question of full or reduced features set is often discussed. Furthermore, for acted and non-acted emotions, reduced and full features sets differences are not as big [4].

4. CLASSIFICATION ALGORITHM

A number of factors motivate the consideration of diverse classifications methods: tolerance to high dimensionality, capability of exploiting sparse data, and handling of skewed classes.

Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbour (kNN) classifiers are popular since the very first studies. They turned out to be efficient for acted and non-acted speech but show problems with "the increasing number of features that leads to regions of the feature space where data is very sparse" [1]. Also well known, Support Vector Machines (SVM) is a natural extension of LDCs which provides good generalisation properties even for a large feature vector.

Most used non-linear discriminative classifiers are likely to be Artificial Neural Networks (ANNs) and decision trees. ANNs need greater amount of data and therefore are rarely used for acted-data and even less for non-acted. A decision tree provides a hierarchical classification procedure determined by a sequence of questions. Decision tree are less of a "black box" compared to SVM, since they are based on simple recursive question of the data.

For previous classifications methods (static classifiers), functionals are applied on LDD before their classification. Dynamic classifiers like Dynamic Bayesian Networks, Hidden Markov Models or simple Dynamic Time Warp, permit to bypass this step in the computation by implicitly warping observed feature sequences over time. In the literature [8], we note that static classifiers through functionals often give better performances as emotion is "apparently better modelled on a time-scale above frame-level".

Then, when the spoken content is fixed, the combination of static and dynamic processing may help to improve overall accuracy [9]. Popular approaches to combine classifier are *Bagging*, *Boosting* or, more powerful, meta-classifier that learns 'which classifier to trust when'.

5. CONCLUSIONS

In this paper, several topics have been addressed. First, we discuss about an overview of 64 emotional speech data collection. We have highlighted the fact that most of data speech collections are acted databases but today we need more realistic database. Indeed, we simply do need enough data for modelling and new non-prompted data will not really help the research. Non-acted emotions raise the problem of private rights and labelling. Then we introduced the concept of three-dimensional space of emotions in which we saw that taken into account features of all dimensions, can reach better performance for emotions recognition than using one dimension.

Second, a state of the art of features classification has been presented. Features are separated in the following taxonomy: acoustic, linguistic and functionals. Acoustic and linguistic are extracted first, then can be filtered/transformed before data are processed by classifiers. Functionals are applied on acoustic features to provide a sort of normalisation over time. Today, in the preponderant context of recognizing of non-acted emotions, it is important to use all kind of features, particularly linguistic features which seem to be relevant for spontaneous/real-life speech.

Third, classification methods have been briefly reviewed. Distinction between static - Linear Discriminant Classifiers, k-Nearest Neighbour - and dynamic - Hidden Markov Models, Dynamic Bayesian Networks - classifiers was made. Dynamic classifiers allow to skip the step of applying functionals on LDD by implicitly warping observed feature sequences over time but static classifiers through functionals give often better performances.

Finally, the choice of the speech data collection, the selection of relevant features and the choice of a classifier depend on the type of application we want to build. Today, the challenge is to focus the research on how build non-acted speech data collection, determine which features are the most relevant and find how to use the dimensional space aspect for recognizing emotions in non-acted speech.

6. REFERENCES

[1] Schuller B., Batliner A., Steidl S., Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, February 2011.

[2] Ververidis D., Kotropoulos C. Emotional speech recognition: Resources, features, and methods *Speech Communication*, 48(9):1162–1181, September 2006.

[3] Tato R., Santos R., Kompe R., J.M. Pardo Emotional space improves emotion recognition. In *Proc. ICSLP*, 2002.

[4] Thurid Vogt, Elisabeth Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition *ICME'05*, July 2005.

[5] Hirschberg J. Communication and Prosody: Functional Aspects of Prosody *Speech Communication*, 2002.

[6] Cunningham P., Loughrey J. Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets *Research and development in intelligent systems XXI*, Session 1a:, 33-43, 2005

[7] Kohavi, R., John, G. Wrappers for feature subset selection *Artificial Intelligence*,Vol. 97, No. 1-2, 1997

[8] Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G. Frame vs. turn-level: emotion recognition from speech considering static/dynamic processing *Proc. 2nd Int. Conf. Affective Computing*,Vol. LNCS 4738, 2007

[9] Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G. Combining frame and turn-level information for robust recognition of emotions within speech. *Proc. Interspeech*,Vol. LNCS 4738, 2249-2252, 2007

[10] Schuller B., Valstar M., Eyben F., McKeown G., Cowie R., Pantic M. The First International Audio/Visual Emotion Challenge *ACII 2011*, 2011

[11] Ringeval F., Chetouani M. Exploiting a vowel based approach for acted emotion recognition *HH and HM Interaction*, LNAI 5042, pp 244-255, 2008