Usual voice quality features and glottal features for emotional valence detection

Marie Tahon [1,2], *Gilles Degottex* [3,4], *Laurence Devillers* [1,5]

[1] Human-Machine Communication Department, LIMSI-CNRS, Orsay, France
[2] Computer Science Department, University Paris 11, Orsay, France
[3] Computer Science Department, University of Crete, Greece
[4] FORTH, Institute of Computer Science, Heraklion, Greece
[5] ISHA, University Paris-Sorbonne 4, GEMASS-CNRS, Paris, France
mtahon@limsi.fr, degottex@csd.uoc.gr devil@limsi.fr

Abstract

We focus in this paper on the detection of emotions collected in real-life context. In order to improve our emotional valence detection system, we have tested new voice quality features that are mainly used for speech synthesis or voice transformation: the relaxation coefficient (Rd) and the functions of phase distortion (FPD); but also usual voice quality features. Distributions of voice quality features across speakers, gender, age and emotions are shown over the IDV-HR ecological corpus. Our results conclude that glottal and usual voice quality features are of interest for emotional valence detection even facing diverse kind of voices in ecological situations.

Index Terms: voice quality features, emotional valence detection, shape parameter, real-life data.

1. Introduction

In the field of emotion detection in speech, people mainly use prosodic features such as fundamental frequency, energy and duration. Psychoacoustics [1] have shown that the most important features in emotion perception are speech rate, pitch changes, pitch contours, voice quality, spectral content, energy level and articulation. Montrero [2] have estimated that voice quality were more important than prosodic features to recognize cold anger from joy. A previous study from Gendrot [3] underlined the importance of voice quality in emotional valence perception. In our paper, we study the relevance of some voice quality features for emotional valence detection in ecological data [4].

Previous studies tried to introduce voice quality features in emotion detection systems. Usual voice quality features (jitter, shimmer, unvoiced rate and harmonics-to-noise ratio) are use by Clavel [5] for fear detection. The Interspeech challenge 2009 [6], showed that perturbation such as jitter, shimmer and harmonicity (harmonics-to-noise ratio or spectral tilt) were used by the emotion detection community. Some important fields of research are working with voice quality features: speaking and singing voice analysis, speech synthesis and voice transformation. Sun [7] introduces sub-harmonics and harmonics features able to detect roughness. Wavelets are also supposed to support voice quality information such as vocal effort according to Sturmel [8]. Speech transformation recent work is highly focused on voice quality and expressivity. For example, Beller [9] used formants based features and voice quality (mainly tension in the voice). All those voice quality features must be very important for emotion detection, but we need to evaluate how robust they are when facing a high variability of speakers and of emotions.

Emotion detection in real-life conditions is one of the actual challenges in the community [10]. In this context emotion detection systems must be robust to variables that can not be controlled, such as the type of speaker (his age, his voice quality, his way of speaking), the recording conditions (room acoustics, microphone quality, distance from the microphone to the glottal folds) or the type of emotion that are elicited. In this study, room acoustics do not vary in the whole corpus that is used. Emotion detection is based on speech only without any lexical information.

Section 2 presents the audio corpus we use in our experiments. Voice quality features and their variations with different speakers are presented in section 3. Our emotional valence classification results are summarized in section 4.

2. Description of data

We make use of an ecological corpus collected in an apartment at the Vision Institute (Paris): the IDV-HR corpus in which visually-impaired French people interact with the robot Nao as a robotic domestic assistant [10]. This corpus has been collected in the context of the FUI French ROMEO project (www.projetromeo.com). The project aims to design a social humanoid robot which will be able to assist elderly and disabled people at home in everyday activities. The IDV-HR corpus is what we call ecological, because speakers are not actors; they behave as they would probably do in everyday life. Each participant of the IDV-HR data collection is offered to sit comfortably face to NAO, which is sitting down on a coffee table. The participant is recorded with a high quality lapel-microphone (AKG PT40 Pro Flexx). The sampling frequency is 44kHz and the data have been downsampled to 16kHz. A camera is placed behind the robot and films the upper part of the body of the speaker for further studies.

The corpus IDV-HR features elderly people interacting with the robot. The speaker is asked to play three sessions of five scenarios in which he pictures himself in a situation of waking up in the morning. The robot would come to him to chat about either his health, or the program of the day, etc. The utterances of the robot are spoken through a Text-to-Speech module, and are based on pre-established and fixed sentences. Each of these five scenarios is devoted to a different affective state, which the speaker is asked by the robot to act: well-being, minor illness, depressed, medical distress, happy. Each series of five scenarios differ from the other, by the social attitude of the robot (positive: friendly, empathetic, encouraging, or negative: directive, doubtful, machine-like). The robot is remotely controlled by an experimenter who selects the different social attitudes and the utterances which match the content of the speaker's speech the better. This corpus is quite challenging because the speaker

variability is very important. Each speaker do to not express their emotions the same way, they have different voice characteristics. As the participants are not actors, activation is quite low and emotions are shaded.

22 speakers were recorded in the framework of IDV-HR (11 males and 11 females for a median age of 59). The audio channels were manually segmented and annotated by two expert annotators following a specific annotation scheme [10]. After segmentation, an instance corresponds to one single speaker and is emotionally homogeneous. Three main annotations are used for our experiments: activation, major emotional label and minor emotional label. At least, a total of 6071 instances have been annotated with consensual annotations. Major and minor emotions have been reduced to seven macro-classes: neutral, anger, sadness, fear, boredom, happiness and positive/negative (ambiguous class). After segmentation, instances mean duration is 2.45s (from 0.24s to 5.94s), that generally allows us to have more than one voiced part in each instance. For each instance, we compute acoustic feature sets (prosodic and voice quality). The features are then normalized to speaker, and the instance is classified as "negative" or "positive" according to a model trained a priori.

In the following experiments only a subset of instances are used. As we have seen before, the IDV-HR corpus has been recorded in real-life conditions, so far emotions are not prototypical, they are shaded and activation is mostly low. Even valence detection on prototypical data is a hard task; we have selected instances that have been annotated with a high activation level: all instances annotated as "anger", "joy" (that already have an important activation), but only "negative" instances that have a high activation level. Then negative mainly corresponds to anger and strong boredom instances, whereas positive corresponds to joy and strong satisfaction.

3. Voice quality features

3.1. Usual voice quality features

The Praat tool [11] gives us a large panel of micro-prosodic features. Among them, we have selected the most commonly used. The unvoiced part gives indication on how voiced is the speech signal. Harmonics-to-noise estimates the proportion of noise in the speech signal; it highly depends on the room acoustics. The local jitter and shimmer evaluate the small time variation of fundamental frequency and energy. Each micro-prosodic feature is computed on the whole instance.

3.2. Glottal features

In this paper we want to evaluate the relevance of two glottis features: the relaxation coefficient (Rd) estimated using MSPD2 based method [12] and the Functions of Phase-Distortion (FPD) [13]. The Rd coefficient is one parameter estimated from the Liljencrants-Fant glottal model. The more important the Rd is, the more relax the voice is. The estimation of this parameter depends on the glottal model, this model does not take into account all the range of possibilities the human voice is able to do. Even tough it can be used for speech synthesis or voice transformation, we would like to know if it is robust as well for emotional valence detection. The Rd coefficient varies form 0 (very tense) to 2.5 (relax). Some actual studies try to estimate the Rd parameter on a wider range, 0.3 to 6, and thus allowing the model to be more flexible. A confidence value comes with the Rd coefficient; the more important the confidence is, the more reliable the Rd value is. We have tried to add the mean and standard deviation of this confidence, but the results are slightly worst than the ones we have with Rd only.

The Functions of Phase-Distortion (FPD) characterize mainly the distortion of the phase spectrum around its linear phase component. More precisely, the FPD are also independent of: the duration of the glottal pulse, its excitation amplitude, its position as well as the position of the analysis window and the influence of a minimum-phase component of the speech signal (e.g. the vocal-tract filter). In this study and conversely to [13], we estimate and take into account only the time evolution of the FPD computed from the speech signal without considering any glottal model (up to equation (5) in [13]. This feature seems to be interesting for speech analysis because it does not depend of a model.

Both Rd and FPD are computed on an average of 4 pulses on voiced parts only; on such a time window, the fundamental frequency is assumed to be constant. In order to have one parameter for an entire instance, we choose to take the temporal mean value and the std value for each Rd and FPD parameters. As the Rd value is given with a confidence score, we take the mean Rd values that have a confidence higher than 0.70 only. The FPD features correspond to phases values, we have computed the unwrap mean values, this value is set down to $[-\pi; \pi]$.

3.3. Voice quality features and speaker variation

As the two glottal features Rd and FPD have never been tested on a large amount of spontaneous and ecological data, we need to evaluate how they evolve facing several speakers and emotions.

It seems that elder voices have generally a smaller Rd than younger (see Figure 1). It means the elder voices in our corpus are tenser than the younger. We have chosen the age of 60 to do the distinction between elderly people and younger in order to have all four sets balanced but it is obvious that the age is not reliable to determine if the voice sounds aged or not.

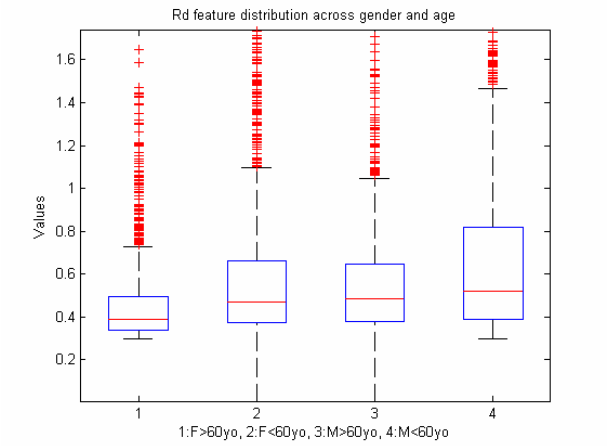

Figure 1: *mean Rd (without speaker normalization) distribution across gender and age (over or under 60) on neutral and emotional instances.*

3.4. Voice quality features and emotion variation

To check if voice quality features vary with emotional valence, we can use an ANOVA test to estimate if voice quality features have a significant impact on valence distribution (see table 1). We have selected four important features that should have some significant impact on emotion detection: mean and std Rd value, unvoiced ratio and the HNR.

HNR seem to be higher for negative instances (median normalized value: 0.371) than for positive instances (-0.919). It means that negative emotions have less spectral noise. The Rd value is significantly more important for positive instances (0.005) than negative instances (-0.032). It means that when feeling a positive sensation, the speaker's voice is more relaxed. The other voice quality features do not have such as significant differences. We are going further in voice quality feature analysis with emotional valence classification.

Table 1. *ANOVA results on speaker normalized voice quality features.*

Voice quality features	p-value
Mean Rd	1.65 e-08
Std Rd	2.88 e-06
Mean FPD	4.51 e-01 (mean)
Punvoiced	2.87 e-07
Jitter	1.71 e-02
Shimmer	2.72 e-01
HNR	<1.00 e-10

4. Emotion detection results

4.1. Acoustic features sets

In this paper we are using different sets of acoustic features computed with the OpenEar (OE) baseline system [14] as a basic set of features. The OpenEar system combines 16 prosodic and spectral features with 12 statistics (minimum, maximum, position of minimum, of maximum, mean, range, standard deviation, linear regression coefficients, kurtosis and skewness). In our experiments we have selected only basic prosodic and cepstral features: fundamental frequency features (11 we removed the minimum of F0 which were always null), energy features (12) and MFCC features (168). We are doing so in order to be able to compare one voice quality feature to the set of acoustic features. Indeed, if the acoustic set has too many features, the influence of one voice quality feature will not be significant.

Table 2. *Features sets*

Basic acoustic features (#features)	Voice quality features (#features)
OE-F0 (11)	**Rd mean and std (2)**
OE-Energy (12)	**FPD1-5 mean (5)**
OE-MFCC (168)	Local jitter (1)
OE-F0+Energy (23)	Local shimmer (1)
	Unvoiced ratio (1)
	HNR (1)

In our experiments we combine one of the four basic acoustic sets (from 11 to 168 features) computed with OpenEar toolkit [14] with voice quality features (5 usual features computed with with Praat [11] and 7 new features coming from a glottal model) as described in table 2. We are also testing without voice quality feature ("alone") or with all features of a same group ("allPraat", "allLibglottis"). New glottal voice quality features are in black in table 2. Once they are computed, each feature is normalized to the mean value of the corresponding speaker.

4.2. Emotion classification

In order to have speaker-independent sets of speakers, we have separated the 22 speakers if IDV-HR corpus in two groups. One experiment consists in training a positive/negative model on a feature set on one group of speakers and testing with the second group and vice-versa. The final detection score is the mean of both tests. All sets are balanced. As we have said before, our data are ecological; it means that everything that comes to the microphone is treated. Since valence detection is a hard task on non-prototypical data, we have selected only instances with a high activation.

Table 3. *train and test sets.*

Valence	Speaker Group n°1	Speaker Group n°2
Negative	210	366
Positive	225	381

4.3. Results

The following experiments are made with features sets that are described in 4.1 and libSVM tool [15] for optimisation of the model parameters. For some experiments it appears that the parameters are very different when testing on group n°1 and on group n°2. Then, both positive and negative classes are classified in one of the two classes. The unweighted average recall can be a good result, but the minimum precision is quite bad. In order to avoid this confusion, our results correspond to the minimum precision on the two classes that have been classified.

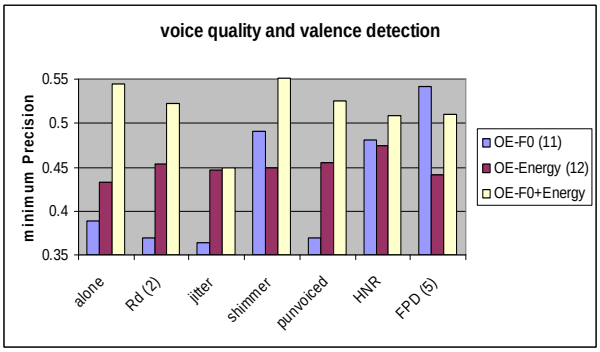

Figure 2: *Positive/Negative with high activation detection using different acoustic sets and voice quality features*

All percentage are reliable at an average of 3.5% for all experiments. This score is a ratio considering the number of instances tested and the classification score. First of all, the MFCC basic acoustic set (see Figure 3) does not lead to a minimum precision over 50% (random guess) in any cases. This is probably due to the fact that MFCCs coefficients are highly related with speaker characteristics, so a speaker-

independent classification task leads to low scores. In figure 2, we can see that the OE-F0+Energy set leads to better results than Energy or F0 alone. With the OE-F0+FPD feature set, the minimum precision is also over the random guess. It could mean that FPD and F0 carry different kind of information surely since FPD are normalized by pulse length. The shimmer associated with OE-F0+Energy gives the best minimum precision. Minimum precision are almost the same with all Praat features and with all Libglottis features (see Figure 3).

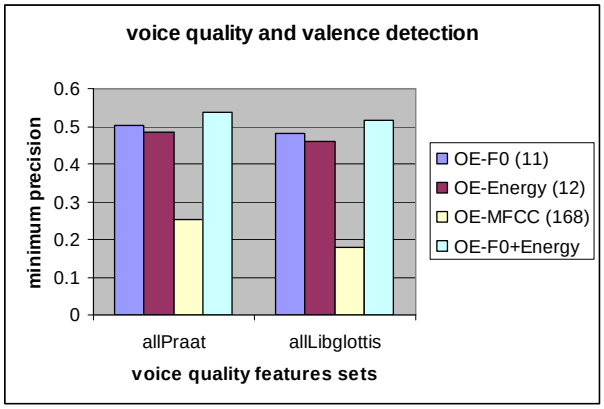

Figure 3: *Comparison of Positive/Negative classification between Praat and Libglottis*

5. Conclusion

Classification results may be biased because the libSVM classifier optimizes the C and gamma parameters on accuracy. As we have seen before, accuracy can be over 50%, but only one class is recognized. Then, both features analysis and classification are important to estimate the interest of voice quality features.

In section 3, we have shown that Rd parameter is highly related to speaker, and more precisely to its gender and age. Among the 6 voice quality features tested, 4 of them seem to be interesting for valence discrimination: mean and std Rd, HNR and unvoiced ratio. In section 4, we have seen that FPD functions associated with F0 features and the shimmer associated with F0 and Energy features are interesting for valence detection. In IDV-HR corpus, MFCCs features do not seem to be reliable for speaker-independent experiments. Jitter and shimmer, do not conclude to good results in our study, we probably need to estimate them on some specific signals (vowels, consonants, phonemes, fixed time window) or as we have done for estimation of Rd and FPD, on a small number of pulses. Estimation of Rd parameter on wider range should lead to a better classification score with Rd features. In order to improve performances of the classification method, we need first, to optimize the model with more data, and secondly, to avoid the libSVM classifier bias we have mentioned before. In this paper we have chosen to compute Rd and FPD features on voiced parts only; further studies will try to get more precise time window (phoneme, syllable, fixed) for high-level features computation (such as Rd, FPD, but also jitter, shimmer, etc.).

As they try to represent the speech signal in very exact way, features developed in speech synthesis or speech transformation support very useful information for emotion detection. Our results show that voice quality features such as Rd parameter and FPD functions might be useful for emotion detection, even facing diverse kind of voices in ecological situations. Further studies will integrate those features to a more complete set to improve emotional valence detection. This introduces a new challenge: is it possible to have features that are as flexible as possible to face the huge variability we have in real-life interactions?

6. Acknowledgements

This work is financed by national funds FUI6 under the French ROMEO project labelled by CAP DIGITAL competitive centre (Paris Region).

7. References

[1] Shilker, T.S., “Analysis of affective expression in speech”, PhD thesis, Cambridge University, Computer Laboratory, chap.5, 2009.

[2] Montero, J.M., Gutiérrez-Arriola, J., Colas, J., Enriquez,E. and Pardo, J.M., « Analysis and modelling of emotional speech in spanish”, ICPhS 1999, San Fransisco, U.S.A., 1999.

[3] Gendrot, C., “Rôle de la qualité de la voix dans la perception des emotions: une etude perceptive et physiologique”, Parole, 13(1) : 1-18, 2004.

[4] Clavel, C., Devillers, L., Richard, G., Vidrascu, I. and Ehreette, T., “Abnormal situations detection and analysis through fear-type acoustic manifestations”, ICASSP 2007, Honolulu, Hawaii, U.S.A., April 2007.

[5] Douglas-Cowie, E., Cowie, E., Schroeder, M., “Research on the expression of emotion is underpinned by databases. Reviewing available resources persuaded us of the need to develop one that prioritised ecological validity”, in SpeechEmotion, 39-44, 2000.

[6] Schuller, B., Batliner, A., Steidl, S. and Seppi, D., “Recognizing realisitc emotions and affect in speech: state of the art and lessons learnt from the first challenge”, Speech Communication, Elsevier, 2011.

[7] Sun, X., “Pitch determination and voice quality analysis using subharmonic to harmonic ratio”, ICASSP 2002, Orlando, Florida, U.S.A., May 2002.

[8] Sturmel, N., d’Allessandro, C., Rigaud, F., “Glottal closure instant detection using lines of maximum amplitudes (LOME) of the wavelet transform”, ICASSP 2009, Taïpeï, Taïwan, April 2009.

[9] Beller, G., “Expresso : Transformation of Expressivity in Speech”, Speech Prosody, Chicago, May 2010..

[10] Tahon, M., Delaborde, A., Devillers, L., “Real-life emotion detection from speech in Human-Robot Interaction: experiments across diverse corpora with child and adult voices”, Interspeech 2011, Firenze, Italy, august 2011.

[11] Boersma, P., and Weenink, D., “Praat: doing phonetics by computer”, from http://www.praat.org/, retrieved May, 2009

[12] Degottex, G., Roebel, A., Rodet, X., “phase minimization for glottal model estimation”, IEEE Transactions on Acoustics Speech and Language Processing, 19(5): 1080-1090, 2011.

[13] Degottex, G., Roebel, A. and Rodet, X., “Function of phase-distortion for glottal model estimation”, In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p 4608-4611, 2011

[14] Schuller, S., Steidl, S., Batliner, A., “The INTERSPEECH 2009 emotion challenge“, in Proc. Of the 10th Interspeech Conference, Brighton, U.K., 2009.

[15] Chang, C.-C.,and Lin, C.-J., "LIBSVM : a library for support vector machines". ACM Transactions on Intelligent Systems and Technology, vol. 2, n°3, pp. 27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.