DISCRIMINATING EXPRESSIVE SPEECH STYLES BY VOICE QUALITY PARAMETERIZATION

Carlos Monzo, Francesc Alías, Ignasi Iriondo, Xavier Gonzalvo, Santiago Planet

GPMM - Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Quatre Camins 2, 08022 Barcelona, Spain
{cmonzo, falias, iriondo, gonzalvo, splanet}@salle.url.edu

ABSTRACT

In this work, the capability of voice quality parameters to discriminate among different expressive speech styles is analyzed. To that effect, the data distribution of these parameters, directly measured from the acoustic speech signal, is used to train a Linear Discriminant Analysis that conducts an automatic classification. As a result, the most relevant voice quality patterns for discriminating expressive speech styles are obtained for a diphone and triphone Spanish speech corpus with five expressive speaking styles: neutral, happy, sad, sensual and aggressive.

Keywords: Voice quality, expressive speech style, discrimination.

1. INTRODUCTION

Automatic speech recognition and text-to-speech (TTS) synthesis are research fields where expressive speech is being used in order to improve the naturalness of human-machine interaction, e.g. in emotion recognition [5] and in voice transformation [6][13]. Voice quality (henceforth VoQ) and prosody parameters are used to represent emotional content of speech [4]. In spite of the fact that VoQ has been less explored than prosody, recent works propose both data to improve the acoustic modeling of expressive speech [4][8], while other studies are denoted to relate perceived speech features to VoQ parameters [2]. In addition, works as [7] deals with the association of phonation type (e.g. whispery voice) and affective speaking. Thus, prosody, used in recent works [8], is not involved in this work to analyze independently VoQ parameterization on expressive speech styles.

The main aim of this work is to explore the effectiveness of using VoQ parameters to discriminate among expressive speech styles. To that effect, we study the relationship between VoQ parameters and expressive styles, by conducting descriptive statistics and Linear Discriminant Analysis (LDA) on an expressive speech corpus in Spanish.

The paper is organized as follows. In section 2 the speech material and the VoQ parameters are presented. In section 3 VoQ parameters distributions across the speech material and automatic classification method are detailed. In section 4, discrimination experiments are reported and discussed. Finally, in section 5, conclusions are presented.

2. SPEECH PARAMETERIZATION

In this section, the speech parameterization process based on VoQ parameters is presented.

2.1. Speech material

This work makes use of a Spanish speech corpus created to be used for conducting TTS synthesis. This corpus was recorded by a female professional speaker and it is composed of five expressive speaking styles: Neutral (N), Happy (H), Sad (Sa), Sensual (Se) and Aggressive (A) (see Table 1).

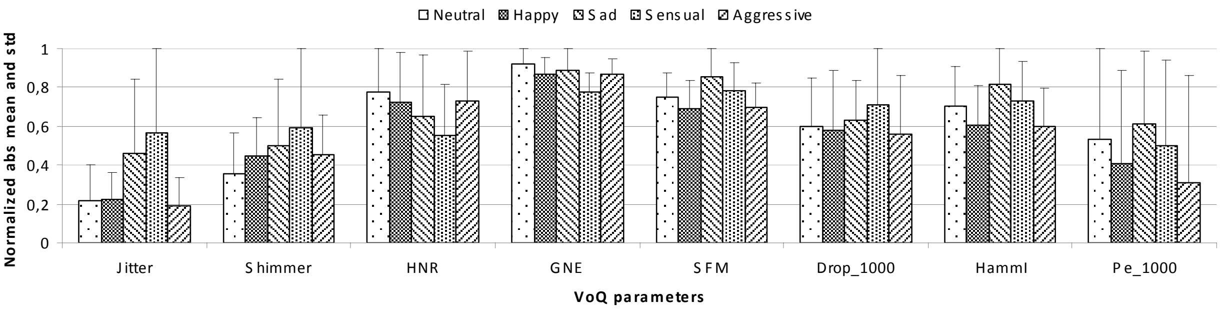
The same phrases were recorded for each expressive speech style (1248 phrases composed of 1 or 2 words). These phrases cover the 30 Spanish phonemes, 831 diphones and 417 triphones. For instance, a pair of phrase samples with their di/triphones of interest (in Spanish SAMPA notation [12]), together with its corresponding English translation, is the following:

- "Pisar" → /pi/ → "to step on"
- "Ruidoso" → /Rwi/ → "noisy"

Table 1: Corpus features per expressive speech style

Style	Duration	Speech Rate	Phonation
Neutral	22'	Fast	Modal
Happy	24.5'	Slow	Mid-harsh
Sad	25.3'	Slow	Mid-whispery
Sensual	30.5'	Very slow	Whispery
Aggressive	24'	Slow	Harsh

Figure 1: Normalized absolute mean and standard deviation of VoQ parameters per expressive speech style



Each expressive speech style is characterized by its own phonation type and speech rate. Therefore, there are different durations although the phrases are the same (see Table 1).

After corpus recording, the speech data were phonetically transcribed and segmented, i.e. the beginning and the end of each phoneme were tagged. Thus, phonemes under test are easily identified in each audio file.

2.2. VoQ selected parameters

The VoQ parameters used in this study are directly estimated from the acoustic speech signal by means of Praat [3], therefore, neither extra hardware nor invasive transducers are required [9]. According to [6] the following VoQ parameters are selected:

- *Jitter* and *Shimmer*: compute the cycle-to-cycle variations of the fundamental period and waveform amplitude, respectively, i.e. describes frequency and amplitude modulation noise. These measurements are called “local” in Praat.
- Parameters describing *additive noise* as *Harmonic-to-Noise Ratio* (HNR) and *Glottal-to-Noise Excitation Ratio* (GNE). GNE is a good alternative to HNR since it is almost independent from jitter and shimmer [10].
- *Spectral Flatness Measure* (SFM), computed as the ratio of the geometric to the arithmetic mean of the spectral energy distribution.
- *Hammarberg Index* (HammI), defined as the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands.
- *Drop-off of spectral energy above 1000Hz* (Drop_1000), a linear approximation of spectral tilt above 1000 Hz, which is calculated using least squares method [1].
- *Relative amount of energy in the high* (above 1000Hz) *versus the low frequency range* of the voice spectrum (Pe_1000).

Although some parameters could seem similar, all of them are necessaries on the conducted experiments to know their behavior.

2.3. Selected data

The VoQ parameters computation is only conducted on the Spanish vowels: /a/, /e/, /i/, /o/ and /u/, as they represent stable voiced zones, which ensure the VoQ parameters goodness.

Moreover, in each analysis for obtaining the selected VoQ parameters, the same number of vowels is used. If an error is reported during the parameters calculation, e.g. too short vowel duration, a code error is generated so as to prune all these cases from the speech material.

3. DISCRIMINATION ANALYSIS

In this section, by means of descriptive statistics and LDA classification, the capability of VoQ parameters to discriminate among expressive speech styles is analyzed.

3.1. Descriptive statistics

Figure 1 shows the normalized absolute means and standard deviations of VoQ parameterization. The main idea is to extract parameters patterns across all expressive styles, and thus, ensuring the viability of conducting a discrimination analysis.

As it can be observed, VoQ parameters present particular data distributions for each expressive style. The most separated distributions are a priori good candidates to be discriminated, although the high dispersion obtained in some of them may be a drawback. Subsequently, an automatic classification process, by means of LDA, helps in the discrimination process.

3.2. LDA for discrimination

LDA analysis is a statistical technique able to classify objects into mutually exclusive and exhaustive groups based on a set of measurable object's features. The main reasons for making use of LDA in this work are: *i)* it has been widely used in VoQ parameterization studies [6][9] and *ii)* there is no need for a long training phase [9], thus, making the process straightforward.

4. EXPERIMENTS

The aim of the conducted analysis is finding the most representative VoQ parameters that discriminate expressive speech styles. Therefore, by means of descriptive statistics and LDA classification of parameterized data (see section 2.3) the discrimination is conducted. The results obtained from automatic LDA classifier are validated by means of t-test on data distributions. Significance level is measured and analyzed among expressive speech styles pair-wise comparison per VoQ parameter.

For each VoQ parameter (input data to the classifier), all expressive speech styles (output classifier classes) are known. By means of 10-fold cross validation (using a random process for data selection) the training and testing information are obtained. The F1 measure [11] is used to assess the classifier performance as it combines both precision and recall into a single metric and favors a balanced performance of the two metrics.

4.1. Automatic expressive style discrimination

After the classifier training, experiments are conducted. Figure 2 shows the LDA performance for discriminating expressive styles through VoQ parameters in terms of F1 measure.

As a result, the F1 measures lower than 0.3 are discarded due to they are not significant, and thus, the best discriminating parameters per expressive style are obtained (see Table 2). In addition, the most relevant VoQ parameters per style relations are shown in Table 3. Both tables complement themselves.

Table 2: The highest discrimination of expressive speech styles by VoQ parameterization

VoQ parameter	Expressive style
Jitter	Se, A
Shimmer	N, Se
HNR	N, Se
GNE	N, Se
SFM	H, Sa
Drop_1000	Se, A
HammI	Sa, A
Pe_1000	Sa, A

Table 3: The most relevant relations among expressive speech styles and VoQ parameters

Expressive style	VoQ parameter
Neutral	Shimmer, HNR, GNE
Happy	SFM
Sad	SFM, HammI, Pe_1000
Sensual	Jitter, Shimmer, HNR, GNE, Drop_1000
Aggressive	Jitter, Drop_1000, HammI, Pe_1000

Notice that all parameters are important to discriminate the expressive speech styles. Results in Table 2 show that 'Sensual' and 'Aggressive' are the most discriminated styles. 'Sensual' style, which usually is difficult to be identified by only using prosodic parameters [8], can be clearly discriminated. Otherwise, 'Happy' style is only discriminated by means of SFM parameter, indicating the low relevance of this expressive style with VoQ parameterization.

Moreover, these results can be also interpreted in terms of phonation type analysis. For instance (see Table 1), 'Sensual' style is characterized by a whispering voice (noisy), whereas 'Neutral' style is characterized by a modal voice, therefore, VoQ parameters as GNE, related to noise measurement, make possible the discrimination between them.

Figure 2: F1 measure for five expressive speech styles using VoQ parameters

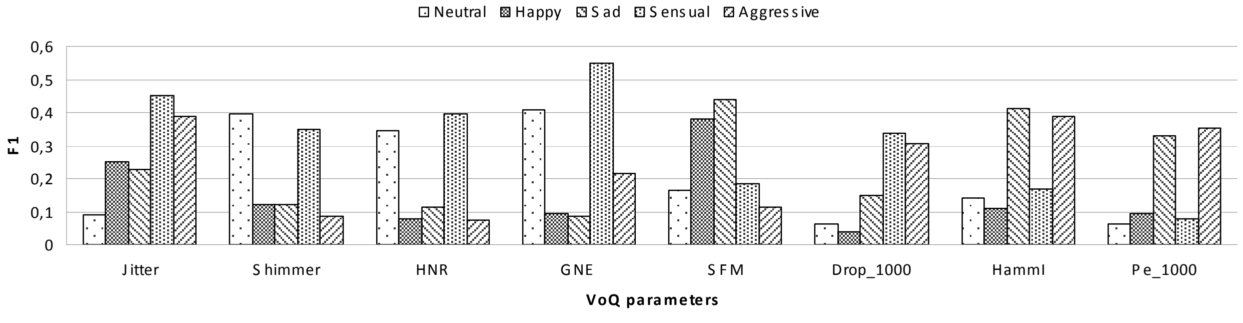


Table 4: Significance level value for expressive speech styles pair-wise comparison per VoQ parameter (threshold for significance level is: * p < 0.05)

	N-H	N-Sa	N-Se	N-A	H-Sa	H-Se	H-A	Sa-Se	Sa-A	Se-A
Jitter	0.25	*	*	*	*	*	*	*	*	*
Shimmer	*	*	*	*	*	*	0.31	*	*	*
HNR	*	*	*	*	*	*	0.63	*	*	*
GNE	*	*	*	*	*	*	0.44	*	*	*
SFM	*	*	*	*	*	*	0.32	*	*	*
Drop_1000	0.09	*	*	*	*	*	0.08	*	*	*
HammI	*	*	*	*	*	*	0.45	*	*	*
Pe_1000	*	*	0.11	*	*	*	*	*	*	*

4.2. Discrimination results validation

Once the classification using LDA is conducted is necessary to validate the obtained results statistically. In Table 4, the significance level (p), calculated on data distribution (see Figure 1) is shown for each expressive speech styles through a pair-wise comparison. Below the threshold value ($p < 0.05$), expressive styles pair-wise comparison is considered significantly different, and thus, making possible the discrimination.

Notice that the clearest discrimination results obtained from LDA analysis (see Table 2) are below the significance level, therefore LDA discrimination among expressive speech styles is validated. However, there is one clear exception between ‘Happy’ and ‘Aggressive’ styles, where significance level for SFM, Drop_1000 and HammI indicate that these VoQ parameters are not useful for discriminating between these styles.

5. CONCLUSIONS

In this work, different voice quality parameters (VoQ), computed from an expressive speech corpus in Spanish, have been analyzed to study their capacity of discrimination among five expressive speech styles. In order to automate this discrimination process, an LDA classifier, based on VoQ parameters, has been used and statistically validated.

The effectiveness of the LDA classifier for expressive speech style discrimination has been statistically validated and discussed. It can be observed that all VoQ parameters are useful to discriminate among the five expressive speech styles of the analyzed corpus. However, we also conclude that prosodic information could be necessary for discriminating between ‘Happy’ and ‘Aggressive’ styles.

There are several practical applications for this work, such as expressive TTS synthesis and emotion recognition. For instance, we could apply voice transformation rules from the LDA classification results and define pattern recognition from VoQ data distributions. In future work, experiments with other automatic classifiers will be conducted.

6. ACKNOWDLEGEMENTS

This work was supported by the SALERO project (IST-FP6-027122) of the European Commission.

7. REFERENCES

[1] Abdi, H. 2003. Least squares. In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks (CA): Sage, 559-561.

[2] Bänziger, T., Scherer, K. 2003. A study of perceived vocal features in emotional speech. In *VOQUAL'03*, Geneva, 169-172.

[3] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International,* 5:9/10, 341-345

[4] Cabral, J., Oliveira, L. 2005. Pitch-synchronous time-scaling for high-frequency excitation regeneration. In *INTERSPEECH,* Lisbon, 1513-1516.

[5] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1): 32-80.

[6] Drioli C., Tisato G., Cosi P., Tesser F. 2003. Emotions and voice quality: experiments with sinusoidal modeling. In *VOQUAL’03*, Geneva, 127-132.

[7] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189-212.

[8] Iriondo, I., Planet S., Socoró, J.C., Alías, F. 2007. Objective and subjective evaluation of an expressive speech corpus. In NOLISP07, 15-18, Paris.

[9] Lugger, M., Yang, B. 2006. Classification of different speaking groups by means of voice quality parameters. *ITG Sprachkommunikation*, Kiel.

[10] Michaelis, D., Gramss, T., Strube H.W. 1997. Glottal to noise excitation ratio - a new measure for describing pathological voices. *Acustica/acta acustica*, 83, 700-706

[11] Sebastiani, F. 2002. Machine learning in automated text Categorization. ACM Computing Surveys, 34(1), 1-47

[12] Spanish SAMPA notation: http://www.phon.ucl.ac.uk/home/sampa/spanish.htm visited 10-Jan-07

[13] Turk, O., Schröder, M., Bozkurt, B., Arslan, L.M. 2005. Voice quality interpolation for emotional text-to-speech synthesis. In *INTERSPEECH*, Lisbon, 797-800.