EXPRESSIVE SPEECH CORPUS VALIDATION BY MAPPING SUBJECTIVE PERCEPTION TO AUTOMATIC CLASSIFICATION BASED ON PROSODY AND VOICE QUALITY[*]

Ignasi Iriondo, Santiago Planet, J. Claudi Socoró, Francesc Alías, Carlos Monzo, Elisa Martínez

GPMM. Enginyeria i Arquitectura La Salle. Universitat Ramon Llull. Barcelona (Spain)
{iriondo,splanet,jclaudi,falias,cmonzo,elisa}@salle.url.edu

ABSTRACT

This paper presents the validation of the expressiveness of an acted corpus produced to be used in speech synthesis, as this kind of emotional speech can be rather lacking in authenticity. The goal is to obtain a system which is able to prune bad utterances from an expressiveness point of view. The results from a previous subjective test are used for the training of a multistage emotional identification system based on statistical features from prosody and voice quality. As a result, a set of utterances is provided to be checked and definitely eliminated if appropriate.

Keywords: Emotional Speech, Speech Synthesis.

1. INTRODUCTION

One of the most important challenges in the study of expressive speech is the development of oral corpora with authentic emotional content. The naturalness will depend on the strategy followed to obtain emotional speech. The main debate is centered on the compromise between authenticity and the degree of control over the recording. In [2, 10], four emotional speech sources are proposed: natural, elicited, stimulated or acted. Speech synthesis is a process of expression (*centred on the listener*) [10], where speech is modelled in order to transmit emotions. Databases for emotional speech synthesis are usually based on acted speech [6], where a professional speaker reads a set of texts (neutral or with emotional content) simulating the desired emotions. More information about emotional databases can be found in [3, 5, 11].

This paper describes the main aspects of the production of an expressive speech corpus in Spanish to be used for synthesis purposes. The main contribution is the mapping between objective and subjective assessment of the emotional content in order to prune bad utterances (i.e with delivered emotion different from the desired one), improving the quality of the speech material and reducing the time of revision.

2. OUR EXPRESSIVE SPEECH DATABASE

We have developed a new expressive oral corpus for Spanish aimed at speech synthesis with a twofold purpose: firstly, to be used in the acoustic modelling (prosody and voice quality) of the emotional speech, and secondly, as a speech unit database for the synthesizer. For the recording, a female professional speaker read texts semantically related to different expressive styles, which were selected from a textual database of advertisements. Based on a study of voice in audio-visual publicity [8], five categories were chosen and then, the most suitable style was assigned to each one: new technologies (neutral-mature), education (joy-elation), cosmetics (style sensual-sweet), automobiles (aggressive-hard) and trips (sad-melancholic). The recorded database has 4638 phrases and it is 5 h 12 min long.

Prosodic features of speech (fundamental frequency, energy, segmental duration and pausing) and voice quality are both related to the vocal expression of emotion [4]. An automatic acoustic analysis of the utterances is performed using information from the previous phonetic segmentation. The analysis of the fundamental frequency (F0) parameters is based on the system described in [1]. In that, unvoiced segments and silences are marked using interpolated values from the neighbouring voiced segments. For energy, speech is processed with 20-ms rectangular windows and 50% of overlap, computing the mean energy in decibels (dB). Moreover, we incorporated rhythm information using the z-score as a means to analyze the temporal structure of the speech and besides considering the following pausing parameters: frequency and duration of pauses.

The voice quality (VoQ) parameters (see [9]) involved in the present study are: i) Jitter and Shimmer (cycle-to-cycle variations of the fundamental period and waveform amplitude respectively); ii) Glottal-to-Noise Excitation Ratio (GNE); iii) Hammarberg Index (HammI), defined as the difference between the maximum energy in the 0-2000 Hz and 2000-5000 Hz frequency bands; and iv) Drop-off of spectral energy above 1000Hz (Do1000).

[*] This work has been partially supported by the European Commission, project SALERO FP6 IST-4-027122-IP.

3. VALIDATION SCHEMA

In this section, the approach for validating the expressiveness of the corpus utterances is presented. Figure 1 shows a block diagram that summarizes the main contribution of this approach. Firstly, a listening test was conducted with almost ten percent of the corpus utterances. From the identification results of this test, a simple decisional system was developed in order to classify the test utterances in two classes: correct and bad. Bad utterances –from an expressiveness point of view– should be removed from the corpus according to subjective criteria. Secondly, we have developed an automatic emotion identification system based on statistical features computed from acoustic parameters of the speech [7]. Three algorithms have been trained and tunned in order to map their behaviour with the subjective criteria shown by the subjective test. Finally, different classifiers were combined by means of a stacking technique [12] in order to improve the results of a single configuration. Once this layer is trained, a list of the candidate utterances to be pruned is generated by running the process to the whole corpus.

3.1. Subjective test

An exhaustive evaluation of the whole corpus (4638 utterances) would be excessively tedious. In order to have a significant sample of subjective perception, 480 utterances (96 per style) were randomly chosen. A forced answer test was designed with the question: *What emotional state do you recognize from the voice of the speaker in this sentence?* The possible answers are the 5 styles of the corpus plus one more option *Don't know / Another* (Dk/A) to avoid biasing the results due to confusing cases. The evaluators were 30 volunteers (19 male and 11 female) of *Enginyeria i Arquitectura La Salle* with quite a heterogeneous profile.

The results of the subjective test show a high percentage of identification (87% in average). The confusion matrix (table 1) shows that sad style (SAD) was the best rated (98.5% in average), followed by sensual (SEN) (87.2%) and neutral (NEU) (86.1%), and finally happy (HAP) (81.9%) and aggressive (AGR) (81.6%). The main confussion happens between AGR and HAP. Moreover, there is also a certain confusion between SEN with SAD and NEU. Dk/A option was seldom used, though it was more present in NEU and SEN than for the other styles.

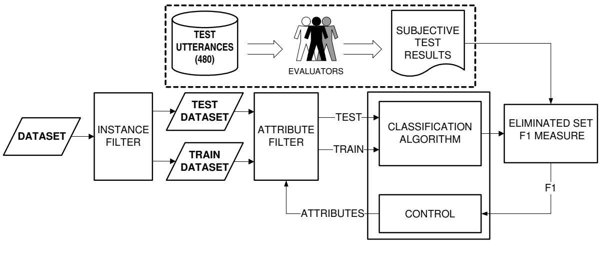

Figure 1: Objective validation adjustment guided by the subjective test results

Table 1: Confusion matrix for the subjective test

	AGR	HAP	NEU	SEN	SAD	Dk/A
AGR	81.6%	15.5%	1.6%	0.1%	0.1%	1.2%
HAP	15.1%	81.9%	1.6%	0.2%	0.1%	1.2%
NEU	5.7%	1.5%	86.1%	3.4%	0.8%	2.5%
SEN	0.0%	0.1%	4.2%	87.2%	6.0%	2.6%
SAD	0.0%	0.0%	0.4%	1.0%	98.5%	0.1%

3.2. Statistical analysis and dataset

In [7], an experiment of automatic emotion identification covering different datasets of statistical prosodic features and algorithms was performed. Initially, we began with a dataset of 464 prosodic attributes per utterance, which was divided into different subsets according to some strategies to reduce its dimensionality. The experiment carried out in the full set of attributes showed almost the same results than in a dataset reduced to 68 parameters. Therefore, we chose this dataset as basis for this work. In this dataset, the prosody of an utterance is represented by the vectors of logF0, energy in dB and normalized durations (z-score). For each sequence, the first derivative is calculated. The result of this work showed that this prosodic dataset was not able to distinguish sensual from sadness, thus, we decided to include VoQ parameters (described in section 2.). For this reason, the five parameters previously described in section 2 were computed over the vowels of the utterance. For all these sequences, the following statistics were obtained: mean, variance, maximum, minimum, range, skew, kurtosis, quartiles, and interquartilic range. Thus, 123 parameters by utterance were calculated, also considering both parameters related to the pausing (see figure 2).

3.3. Supervised classification

Numerous schemes of automatic learning can be used in a task such as classifying the style/emotion from the acoustic analysis of the speech. The objective assessment of expressiveness in our speech

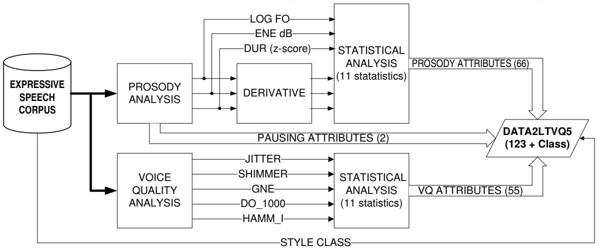

Figure 2: Database generation for automatic classification

corpus is based on [7], where twelve machine learning algorithms were tested with different datasets. All the experiments were carried out using Weka software [12] by means of ten-fold cross-validation. Very high identification rates were achieved. For instance, the average behaviour for all datasets showed that, SMO (Support Vector Machine of Weka) obtained the best results ($\sim$ 97%) followed by NB (Naïve-Bayes) with 94.6% and J48 (Weka Decision Tree based on C4.5) with 93.5%. With these results, we could conclude that, in general, the styles of the acted speech corpus can be clearly distinghished automatically, although the results from the subjective test showed there is certain confusion among styles which was not detected by the system. Therefore, we considered it was necessary to go one step further by developing a method to validate each utterance following subjective criteria instead of a simple automatic classification from the space of attributes.

3.4. Subjective-based attribute selection

In this work, a method for attribute selection has been developed in order to determine the subset of attributes that better maps the subjective test results. As previously mentioned, the original set has 123 attributes per utterance and therefore, an exhaustive search of subsets is not practical. Then, a greedy search procedure has been used, which is guaranteed to find a locally optimal set of attributes [12]. We have chosen a combined method of *Forward Selection* (FW) process, which starts without no attribute and adds them one at a time, and *Backward Elimination* (BW) which deletes the least significant attributes, one at a time. The number of forward and backward steps can be adjusted (e.g. 3FW+1BW). For each iteration, the classifier is trained with the 4158 utterances that do not belong to the test set described in section 3.2. The test utterances are automatically classified according to one of the 5 styles. Moreover, the subjective punctuation of these utterances is also available. The wrongly classified cases take part in the assessment process of the involved subset of attributes. The novelty of this process is the use of a subjective-based measure to evaluate the expected performance of the attributes in each iteration. The used measure is the F1 score computed from the *precision* and the *recall* of the wrongly classified utterances compared with the subset of utterances rejected during the subjective test. For the test subset, an utterance is considered not correctly performed by the professional speaker if it had an identification percentage lower than 50% or Dk/A percentage larger than 12%. There are 33 utterances out of 480 that satisfy at least one rule (6.88% of reduction due to lack of the adequate expressiveness).

3.5. Stacking

Different algorithms of automatic classification and different attribute selection strategies have been tested, obtaining some solutions with high precision and others with high recall. Experience has shown that combining the predictions from multiple methods often yields more accurate predictions, which can be derived from any one of them [12]. In the stacking strategy, the predictions from different classifiers are used as an input to a meta-learner, which attempts to combine these predictions to create a final best predicted classification. In our case, the final classification consists of correct or bad utterances. Therefore, the set of binary outputs from each classifier is the input to this meta-learner. The most simple solution is implementing a voting schema although other simple learners such as tree-based or rule-based classifiers can achieve better results.

4. EXPERIMENTS AND RESULTS

For the single classifiers, two improvements have been incorporated with respect to the baseline [7]. On one hand, the dataset has been completed with VoQ parameters to reduce the misclassification of certain styles (e.g. sensual). On the other hand, a bidirectional attribute reduction has been used in order to avoid getting stuck in poor local maxima. Compared with the baseline, both improvements suppose a relative increase higher than 20% in terms of F1 measure (see table 2). Figure 3 shows the evolution of the maximum of F1 measure according to the best subset of attributes in each iteration of the 3 forward - 1 backward strategy.

Table 2: F1 for the optimum SMO, J48 and NB adding VoQ and 3FW+1BW to the baseline [7]

Algorithm	Baseline F1	VoQ F1	FW-BW F1
SMO	0.50	0.59	0.61
J48	0.45	0.53	0.58
NB	0.42	0.48	0.54

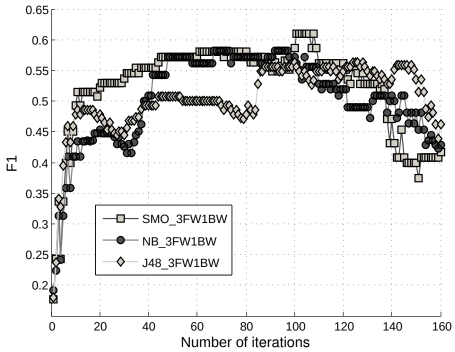

Figure 3: Maximum F1 values per iteration for 3FW+1BW attribute selection.

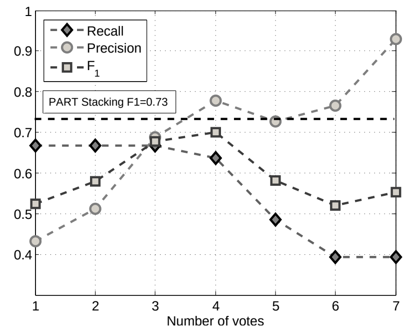

Figure 4: F1, recall and precision values according to the number of votes

For the stacking implementation, initially we have tested a simple voting schema with seven classifiers obtained from different combinations of the algorithm (SMO, J48, NB) and attribute selection methods. We have observed that the AGR style is rapidly punished when the number of votes is increased. Therefore, this class has been weighted twice the other styles. The highest $F1$ measure is 0.7, which is achieved with 4 votes, followed by 3 votes, which yields a closer recall and precision values (figure 4). The highest single result 0.61 is considerably improved (see table 2). Moreover we have trained a rule based classifier based on PART [12] that slightly improves this result obtaining $F1 = 0.73$. The classifiers used by PART are C1=SMO(3fw-1bw), C2=J48(3fw-1bw), C3=J48(4fw-1bw) and C4=NB(3fw-1bw). Algorithm 1 shows the final rules where 0 means correctly classified and 1 misclassified. It is worth pointing out that AGR style is processed specifically by the second rule. Finally, both stacking strategies (voting and PART) have been applied to the whole corpus obtaining different subsets of sentences to be eliminated. The highest vote will provide high precision while decreasing vote will increase the recall. Figure 5 shows the percentage of eliminated utterances detailed per style. Notice that, in addition, "Voting-4" and "PART" present similar global results to subjective evaluation (see table 1).

Algorithm 1 PART decision list for stacking

```
C1 = 0 and C2 = 0 and C3 = 0 and C4 = 0: CORRECT (408/10)
C1 = 0 and Style = AGR and C2 = 1: BAD (7/2)
C1 = 0: CORRECT (25)
C3 = 1: BAD (18/3)
C2 = 0 and C4 = 0: CORRECT (4/1)
C2 = 0: BAD (2)
: CORRECT
```

5. CONCLUSIONS

We have proposed a multilayer automatic classifier that provides a set of utterances with poor or bad expressiveness with respect to the desired one in order to be eliminated. The system has been tuned to map subjective perception, obtained through a previous listening test. In the next step, we will study the suitability of the proposed method by performing acoustic modelling of the resulting emotional speech after pruning according to the results from the whole corpus.

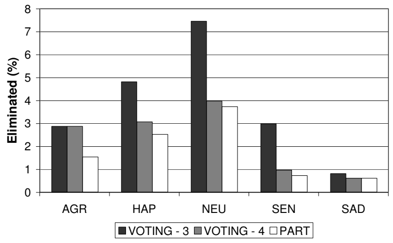

Figure 5: Eliminated utterances per style

6. REFERENCES

[1] Alías, F., Monzo, C., Socoró, J. C. 2006. A pitch marks filtering algorithm based on restricted dynamic programming. *International Conference on Spoken Language Processing.* Pittsburgh (USA).

[2] Campbell, N. 2000. Databases of emotional speech. *Proceedings of the ISCA Workshop on Speech and Emotion,* Northern Ireland (UK). 34–38.

[3] Cowie, R., Douglas-Cowie, E., Cox, C. 2005. Beyond emotion archetypes: databases for emotion modeling using neural networks. *Neural Networks.*

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

[5] Devillers, L., Vidrascu, L., Lamel, L. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18.

[6] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P. 2003. Emotional speech: towards a new generation of databases. *Speech Communication* 40.

[7] Iriondo, I., Planet, S., Alías, F., Socoró, J., Martínez, E. 2007. Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. *9th Int. Work-Conference on Artificial Neural Networks (IWANN).* Donostia, Spain.

[8] Montoya, N. 1998. El papel de la voz en la publicidad audiovisual dirigida a los niños. *Zer. Revista de estudios de comunicación* (4), 161–177.

[9] Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., Planet, S. August 2007. Discriminating expressive speech styles by voice quality parameterization. *Proc. of ICPhS'07.* Saarbrücken (Germany).

[10] Schröder, M. 2004. *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis.* PhD thesis PHONUS 7, Saarland University.

[11] Ververidis, D., Kotropoulos, C. Sept. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication* 48(9).

[12] Witten, I., Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann, 2nd edition.