Emulating Subjective Criteria in Corpus Validation

Ignasi Iriondo
Universitat Ramon Llull, Spain

Santiago Planet
Universitat Ramon Llull, Spain

Francesc Alías
Universitat Ramon Llull, Spain

Joan-Claudi Socoró
Universitat Ramon Llull, Spain

Elisa Martínez
Universitat Ramon Llull, Spain

INTRODUCTION

The use of speech in human-machine interaction is increasing as the computer interfaces are becoming more complex but also more useable. These interfaces make use of the information obtained from the user through the analysis of different modalities and show a specific answer by means of different media. The origin of the multimodal systems can be found in its precursor, the "Put-That-There" system (Bolt, 1980), an application operated by speech and gesture recognition.

The use of speech as one of these modalities to get orders from users and to provide some oral information makes the human-machine communication more natural. There is a growing number of applications that use speech-to-text conversion and animated characters with speech synthesis.

One way to improve the naturalness of these interfaces is the incorporation of the recognition of user's emotional states (Campbell, 2000). This point generally requires the creation of speech databases showing authentic emotional content allowing robust analysis. Cowie, Douglas-Cowie & Cox (2005) present some databases showing an increase in multimodal databases, and Ververidis & Kotropoulos (2006) describe 64 databases and their application. When creating this kind of databases the main arising problem is the naturalness of the locutions, which directly depends on the method used in the recordings, assuming that they must be controlled without interfering the authenticity of the locutions. Campbell (2000) and Schröder (2004) propose four different sources for obtaining emotional speech, ordered from less control but more authenticity to more control but less authenticity: i) natural occurrences, ii) provocation of authentic emotions in laboratory conditions, iii) stimulated emotions by means of prepared texts, and iv) acted speech reading the same texts with different emotional states, usually performed by actors.

On the one hand, corpora designed to synthesize emotional speech are based on studies centred on the listener, following the distinction made by Schröder (2004), because they model the speech parameters in order to transmit a specific emotion. On the other hand, emotion recognition implies studies centred on the speaker, because they are related to the speaker emotional state and the parameters of the speech. The validation of a corpus used for synthesis involves both kinds of studies: the former since it will be used for synthesis and the latter since recognition is needed to evaluate its content. The best validation system is the selection of the valid utterances[1] of the corpus by human listeners. However, the big size of a corpus makes this process unaffordable.

BACKGROUND

Emotion recognition has been an interesting research field in human-machine interaction for long, as can be

observed in Cowie et al. (2001). Some studies have been carried out to observe the influence of emotion in speech signals like the work presented by Rodríguez et al. (1999), but more recently, due the increasing power of modern computers that allows the analysis of huge amount of data in relatively small time lapses, machine learning techniques have been used to recognise emotions automatically by using labelled expressive speech corpora. Most of these studies have been centred on few algorithms and little sets of parameters.

However, recent works have performed more exhaustive experiments testing different machine learning techniques and datasets, as the described by Oudeyer (2003). All this kind of studies had the goal of achieving the best possible recognition rate obtaining, in many cases, better results than those obtained in subjective tests ((Oudeyer, 2003), (Planet, Morán & Formiga, 2006), (Iriondo, Planet, Socoró & Alías, 2007)). Nevertheless, many differences can be found when analyzing the results obtained from objective and subjective classifications and, to our knowledge, there are not studies with the goal of emulating these subjective criteria before those carried out by Iriondo, Planet, Alías, Socoró & Martínez (2007).

VALIDATION OF AN EXPRESSIVE SPEECH CORPUS BY MAPPING SUBJECTIVE CRITERIA

The creation of a speech corpus with authentic emotional content is one of the most important challenges in the study of expressive speech. Once the corpus is recorded, a validation process is required to prune those utterances that show distinct emotion to their label. This article is based on the work exposed by Iriondo, Planet, Alías, Socoró & Martínez (2007) and presents the production of an expressive speech corpus in Spanish with the goal of being used in a synthesis system, validating it by pruning automatically "bad" utterances emulating the criteria of human listeners.

The Production of the Corpus

The recording of the corpus has been carried out by a female professional speaker. There is a high consensus in the scientific community for obtaining emotional speech by means of this strategy for synthesis purposes (Cowie et al., 2005), although other authors argue in favor of constructing enormous corpora gathered from recordings of the daily life (Campbell, 2005). For the design of texts semantically related to different expressive styles, we have made use of an existing textual database of advertisements extracted from newspapers and magazines. Based on a study of the voice in the audio-visual publicity (Montoya, 1998), five categories of the textual corpus have been chosen and the most suitable emotion/style has been assigned to them: New technologies (neutral-mature), education (joy-elation), cosmetic (style sensual-sweet), automobiles (aggressive-hard) and trips (sad-melancholic). The recorded database has 4638 sentences and it is 5 hours 12 minutes long.

From these categories, a set of sentences has been chosen by means of a greedy algorithm (François & Boëffard, 2002) that has allowed us to select phonetically balanced sentences. In addition to looking for a phonetic balance, phrases that contain foreign words and abbreviations have been discarded because they difficult the automatic process of phonetic transcription and labeling.

The corpus has been segmented in phrases and then in phonemes by means of a semiautomatic process based on a forced alignment with Hidden Markov Models.

Acoustic Analysis

Cowie et al. (2001) show how prosodic features of speech (fundamental frequency (F0), energy, duration of phones, and frequency of pauses) are related to vocal expression of emotion. The analysis of F0 performed in this work is based on the result of the pitch marks algorithm described by Alías, Monzo & Socoró (2006). This system can assign marks over the whole signal, interpolating values from the neighbour phonemes in unvoiced segments and silences. Energy is measured with 20 ms rectangular windows and 50% of overlap, computing the mean energy in decibels (dB) every 10 ms. Also, rhythm parameters have been incorporated using the z-score as a means to analyze the temporal structure of speech (Schweitzer & Möbius, 2003). Moreover, for each utterance two parameters relating the number of pauses per time unit and the percentage of silence respect to the total time are considered.