IrcamCorpusTools: an extensible platform for speech corpora exploitation

Christophe Veaux, Gegory Beller, Xavier Rodet

IRCAM – Institut de Recherche et Coordination Acoustique Musique
1 place Igor Stravinsky, 75004 Paris
E-mail: veaux@ircam.fr, beller@ircam.fr, rod@ircam.fr

Abstract

Corpus based methods are increasingly used for speech technology applications and for the development of theoretical or computer models of spoken languages. These usages range from unit selection speech synthesis to statistical modeling of speech phenomena like prosody or expressivity. In all cases, these usages require a wide range of tools for corpus creation, labeling, symbolic and acoustic analysis, storage and query. However, if a variety of tools exists for each of these individual tasks, they are rarely integrated into a single platform made available to a large community of researchers. In this paper, we propose IrcamCorpusTools, an open and easily extensible platform for analysis, query and visualization of speech corpora. It is already used for unit selection speech synthesis, for prosody and expressivity studies, and to exploit various corpora of spoken French or other languages.

1. Introduction

Corpus based methods are increasingly used for speech technology applications and for the development of theoretical or computer models of spoken languages. These usages range from unit selection speech synthesis (Hunt, 1996) to statistical modeling of speech phenomena like prosody or expressivity (Beller, 2006). In all cases, these usages require a wide range of tools for corpus creation, labeling, symbolic and acoustic analysis, storage and query. However, if a variety of tools exists for each of these individual tasks, they are rarely integrated into a single platform made available to a large community of researchers. In (Oostdijk, 2000), an exploitation environment has been proposed for a specifically structured corpus but does not extend to larger corpora frameworks. A standardization effort has been made in (Gut, 2004) in order to use a same query tool for various speech corpora.

In this paper, we propose IrcamCorpusTools, an open and easily extensible platform based on the Matlab/Octave language for analysis, query and visualization of speech corpora. It is already used for unit selection speech synthesis, for prosody and expressivity studies, and to exploit various corpora of spoken French or other languages.

2. System overview

A general overview of the IrcamCorpusTools platform is illustrated in figure 1. This platform has two main functionalities. The first one is to create databases of speech units from recorded speech corpora. The second functionality is the exploitation of the databases.

The creation of a database takes as inputs a recorded speech signal and its associated orthographic text. In a first step, the symbolic information derived from the text is synchronized with the acoustic signal. It provides a

segmentation of the speech into several types of units corresponding to different representation levels of the speech process. Then, these units are characterized both in the acoustic and symbolic domains, and stored into a database together with their structural relationships.

An effective interface in Matlab/Octave allows database querying and graphical exploration.

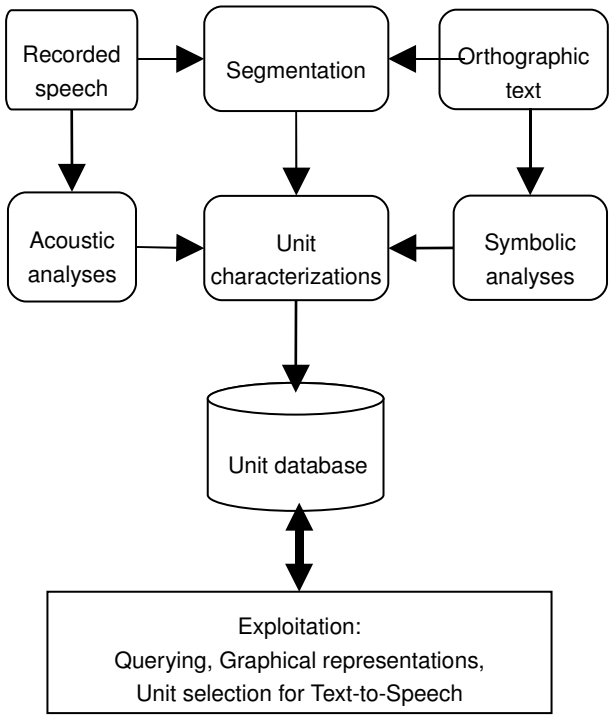

Figure 1: Overview of the IrcamCorpusTools platform

3. Creation of unit databases

3.1 Phonetic and contextual coverage

A set of tools has been developed to compute the phonetic coverage of a corpus as well as its phonetic, lexical and syntactic context coverage. Other tools allow for the construction of a corpus text so as to satisfy coverage requirements such as: all the phones, diphones, etc., in various contexts.

3.2 Segmentation

The first step of a database creation is the automatic segmentation of each speech utterance of a recorded corpus into variable length units: phones, syllable, word tokens, prosodic groups, etc.

Phone segmentation is performed by IrcamAlign, an HMM-based program (Lanchantin, 2008) which takes into account the multiple possible pronunciations of the orthographic text. It also provides the temporal boundaries of each word token. The boundaries of prosodic groups are given by the pauses inserted in the aligned phone sequence. Finally, syllable boundaries are determined by applying a set of syllabification rules to the phone sequence within each prosodic group.

Figure 2 shows some of these different segmentation levels superimposed over the entire speech signal and its fundamental frequency.

3.3 Unit characterization

All segmented units are characterized further by performing both acoustic and symbolic analyses.

3.3.1. Acoustic characterization

A large variety of acoustic features is calculated on the speech signal, such as fundamental frequency, loudness or syllable rate. Then, the temporal evolution of the acoustic features over each unit segment is modeled by a set of single-valued measures:

- arithmetic and geometric mean, standard deviation of the feature over unit segment
- minimum, maximum, and range slope, giving the rough direction of the feature movement, and curvature (from 2nd order polynomial approximation) within unit segment
- value and curve slope at start and end of the unit
- temporal center of gravity/anti-gravity, giving the location of the most important elevation or depression in the feature curve and the first 4 order temporal moments
- 5 bands Fourier spectrum of the feature and their first 4th order moments, which characterize the modulations of the feature over the unit segment

3.3.2. Symbolic characterization

The segmentation step provides phonetically labeled units. Further symbolic labeling is performed automatically such as:

- part-of-speech tags using a French HMM-based POS tagger (Bechet, 2001)

- syllable accentuation, which is predicted from the acoustic domain (Obin, 2008)

Other symbolic labeling can be done manually using annotating tools like (Sjlander, 2000) which have been integrated into our platform.

3.4 Unit relationships

In our data model, we manage two types of structural relationships between units:

- Sequential relationships between units of a same level
- Hierarchical relationships between units of different levels

The sequential relationships are simply accessed by recursive application of operators '*next*' and '*previous*' within the temporal sequence of units.

Hierarchical relationships are represented by a set of trees since a given unit can belong to several hierarchies. For example, phone units belong to a hierarchy that relates them to word tokens, and they also belong to a hierarchy that relates phone units to syllables.

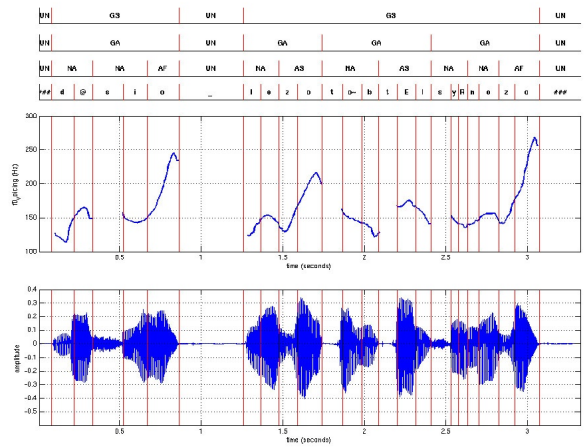

Figure 2: Different segmentation levels superimposed on the fundamental frequency and speech signal

4 Exploitation of a database

A query language has been developed to allow for the exploitation of databases. It relies on the power and the flexibility of the interpreted Matlab/Octave language. The various possibilities offered by this combination are:

- query of units and unit relationships, acoustic features, characteristic values, and labels
- statistical analyses
- graphic display
- export in various formats among which XML.

It is also possible to combine, in a same query, information from different levels of segmentations.

As an example of the power of the interface language, let us show typical commands of this language as given to Matlab:

units = getunits(corpus,{'syllable_accent','is','FinalAccent'});

returns all the accentuated units. Then the command:

histogram(duration(units));

displays the histogram of the duration of these syllable units.

Execution of these two commands for both non-accentuated and accentuated syllables over a database of 8000 neutral French utterances of a male speaker yields to figure 3. It shows that the mean of the accentuated syllable's durations is approximately the double of the mean of the non accentuated syllable's durations.

The database can be also browsed with a graphical database explorer that allows users to visualize all data and play units. Entire utterances can also be accessed through the query language.

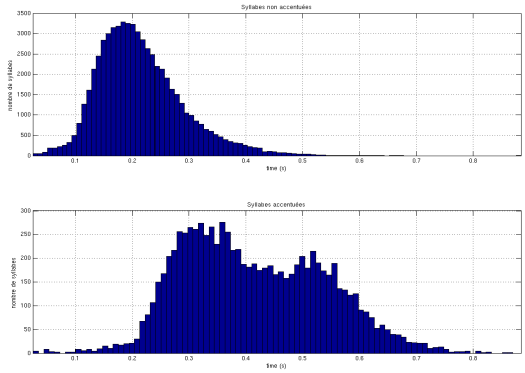

Figure 3: Histograms of syllable duration for the non-accented syllables (top) and the accented syllables (bottom)

5 Easy, extensible and cooperative platform

Since IrcamCorpusTools relies largely on Matlab/Octave, it is easy to install and use for the community of speech

researchers. Extensions and compatibility with various formats can be quickly implemented with simple Matlab/Octave routines. It also has a powerful interactive or batch working environment and excellent debugging facilities. One of the goals of this work is to distribute IrcamCorpusTools and to favor extensions by a community of users with new acoustic and symbolic analysis methods, poweful query constructions, displays and databases. It will also encourage researchers to cooperate, share and compare databases.

6 Applications

The proposed platform is already used for several applications:

- Speech synthesis by unit selection, which uses a corpus of 7 hours of neutral speech from a male speaker.
- Statistical analysis of expressivity, which uses a corpus of approximately 2 hours of french speech. It is composed of neutral and expressive utterances pronounced by a French actors. The corpus is composed of a set of 26 sentences of variable length. Each sentence was pronounced with the following expressivities: neutral, neutral question, angriness, happiness, sadness, boreness, disgust, indignation, positive and negative surprises.
- Statistical analysis of prosody, which uses a corpus of 500 utterances of natural speech from a male speaker. Manual annotation of the syllable accents was performed for each recorded sentence.

7 Conclusion and future works

In this paper, we have presented IrcamCorpusTools, a platform for speech corpora creation and exploitation. It is based on the Matlab/Octave language for analysis, query and visualization which makes it easily to extend. It already integrates several acoustic analyses, automatic labeling tools and interface with standard annotation tools. It supports queries on mutli-level units that make it a powerful tool for statistical exploration of databases.

Several development tasks are being carried out. One of these is the integration of high level symbolic analyses (syntactic and linguistic). One of the main goal is to distribute it and to favor extensions by a community of researchers, especially linguist and phoneticians.

8 Bibliography

A.J. Hunt and A.W. Black, (1996) "Unit selection in a concaenative speech synthesis system using a large speech database," in *Proc. of ICASSP*, Atlanta, GA, pp. 373-376.

G. Beller, D. Schwarz, T. Hueber, X. Rodet, (2006) "Speech Rates in French Expressive Speech," in *Speech Prosody*, Dresden.

N. Oostdijk, (2000) "The Spoken Dutch Corpus: Overview and first evaluation," in *Proc. of LREC*, pp. 887-893.

U. Gut, J-T. Milde, H. Voormann, U. Heid, (2004) "Querying Annotated Speech Corpora," in *Speech Prosody*, Nara, Japan.

P. Lanchantin, A.C. Morris, X. Rodet, C. Veaux, (2008) "Automatic Phoneme Segmentation with relaxed textual constraints," in *Proc. of LREC 2008.*

F. Béchet, (2001) "Lia_phon : un système complet de phonétisation de texts," in *TAL*, vol. 42, no. 1, pp. 47-68.

N. Obin, X. Rodet, A. Lacheret-Dujour, (2008) "French Prominence: A Probabilistic Framework," in *Proc. of ICASSP 2008.*

K. Sjlander and J. Beskow, (2000) "WaveSurfer – an open source speech tool," in *Proc. ICSLP*, Beijing, China.