WinPitch Corpus

A text to Speech Alignment Tool for Multimodal Corpora

Philippe Martin

UFRL Université Paris 7 Denis Diderot
92, Avenue de France, 75013 Paris, France
philippe.martin@linguist.jussieu.fr

Abstract

WinPitch Corpus is an innovative software program for computer-aided alignment of large corpora. It provides a method for easy and precise selection of alignment units, ranging from syllable to whole sentences in a hierarchical storing system of aligned data. The method is based on the ability to link visually and select with a mouse click a text segment with the perception of the corresponding speech sound played back at slower speech. Clicking on a text segment generates bidirectional speech-text pointers defining the alignment. This method has the advantage on emerging automatic processes to be effective even for poor quality speech recordings, or in case of speakers' voice overlap. A recent version of the software handles multimedia files and is capable to display the corresponding video streams at slower speed.

Introduction

Large spontaneous speech corpora are becoming essential for the continuing development of fundamental research in linguistics and language engineering. Commercially available automatic recognition software are usually delivering limited performance for spontaneous speech, due to the use, at the higher stages of the recognition, of syntactic models built from read speech rather than from spontaneous speech data. By the same token, speech synthesis from text would strongly benefit from the emergence of prosodic models inferred from real life data. A better understanding of prosodic interactions linked to spontaneous conversations would also bring spectacular improvement to second language teaching, which is still based for a great part on read material.

Although put a speech corpus together seems to be a simple task per se, involving speech to text transcription and alignment, this ceases to be the case when the volume of data becomes large. Problems inherent to transcription, even for well recorded speech data, are not trivial (Blanche-Benveniste, 2002), and the "manual" transcription and alignment of just one hour of data becomes quickly cost prohibitive, even with the use of modern signal editing software. The development of adequate and user friendly text to speech alignment tools is thus essential for the elaboration of large speech corpora.

Text to speech alignment

Text to speech alignment establishes a bi-univocal relationship between units of speech and units of text. In its simplest implementation, each unit of text (be syllable, word, syntagm or sentence) receives a time index corresponding to the time position in the sound file. When this process is achieved, an operator can select an aligned unit of text and listen to the corresponding speech segment. Acoustical analysis of the speech sound, such as melodic curve and spectrogram, can also be displayed at the same time. Conversely, the selection of a speech segment will highlight the corresponding segment of text, in its orthographic or phonetic transcription.

Spectrographic alignment

Most students in experimental phonetics have some training in spectrographic reading, and are capable of segmenting speech sounds accurately with the visual cues displayed by acoustical analysis. For instance, fricative consonants and stops present easily recognizable graphic features, as do oral vowels. Although other occurrences such as nasals followed by vowels may be harder to segment, the overall process generally leads to a high quality, although labor intensive, segmentation.

In reality, units of text and units of speech cannot correspond exactly, as phonetic and phonological units are defined by human perception on the axis of continuous articulatory transitions, whereas speech signal segments are defined as acoustic entities. Alignment and segmentation can therefore be only approximations, and the physical time limits of speech segments must be positioned somewhere during articulatory transitions of speech sounds.

Nevertheless, transitions between speech sounds are used in some automatic segmentation algorithms, such as (Cosi, 1997). These processes utilize the spectral discontinuities on the time axis, and give acceptable results in otherwise visually clear cases (and may fail for a sequence such as nasal consonant followed by a vowel).

As in all processes of acoustical analysis of speech, reliability of this approach relies on the validity of the hypotheses implied in the process, the most important one assuming the presence of only one sound source in the signal. Background noise and other speech sources will lead to disappointing results in the segmentation.

Automatic alignment with hidden Markov models

Well spread automatic or semi-automatic methods for text to speech alignment utilize algorithms used for speech recognition (often based on parameters obtained by a

Hidden Markov Model applied to the speech data). This approach is a subset of the general speech recognition problem, since the text as already known. The limits of speech segments are then found from a phonetic or orthographic transcription (Talin and Wightman 1994, Fohr, Mari, et Haton 1996).

Although attractive, systems based on automatic speech recognition suffer from the same limitations as speech recognition itself: somewhat high error rate (15% to 20%) without the use of a syntactic model at a higher stage of the process, and the difficulty to train the system with the speaker voices (which must be made from samples of the corpus itself). Again, good results are to be obtained only if the speech signal presents a high signal to noise ratio, and if the voices to align do not differ too much from the models used to train the algorithm. Overlapping speech constitutes of course a very difficult case for these systems.

Automatic alignment by synthesis

Another automatic method of alignment proceeds by comparing the spectral variations of the signal along time with another speech signal, generated by a speech synthesizer fed by the text to align (Malfrère and Dutoit, 2000). The advantage here stems from the fact that it is easier to align successive spectra on two distinct time scale (by dynamic time warping) than segment sounds from automatic recognition of segments.

However the limits of this approach are similar to those of the use of HMM: poor signal to noise level, deviant characteristics of speaker's voices (compared to the models used in the synthesis process) and speech sources overlapping constitute difficult problems for this process.

Limits of automatic alignment

In summary, automatic text to speech alignment processes present the following recurrent limitations:

1. Their performance depends on speaker's voice characteristics, which cannot be too different from the models implied in these methods;

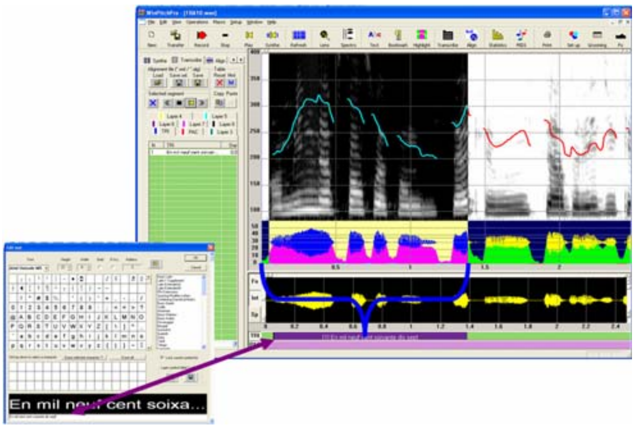
2. The recording signal to noise ratio must be high enough to reduce the error rate to an acceptable level. The radio and TV broadcasts generally meet this requirement, but spontaneous speech recordings made in various public environments (street, public transportation, etc.) present hardly these characteristics. Echo in the speech signal is another perturbing factor;

3. Speaker's voices overlapping, frequently found in spontaneous dialogs constitutes an aggravating factor.

All these considerations seem to indicate that a human operator is required to obtain a reliable text to speech alignment. All the problems mentioned above are then transferred to the operator, who, with appropriate and well ergonomically designed tools, should performed better than by correcting manually the errors made by an automatic system of alignment.

Alignment and transcription

Text to speech alignment can be executed in two modes, depending if the text preexists or not. In the first mode, the text must be created, and the operator proceeds by selecting segments of speech in sequences (which can be played back at reduced speed to enhance intelligibility) and type the corresponding text perceived, either orthographically or in phonetic transcription (WinPitch Corpus allows the use of any font defined in Unicode, the text can be entered directly from the keyboard or through a user defined character map). During this process, a database is automatically updated, and saved directly in Excel® or XML format.



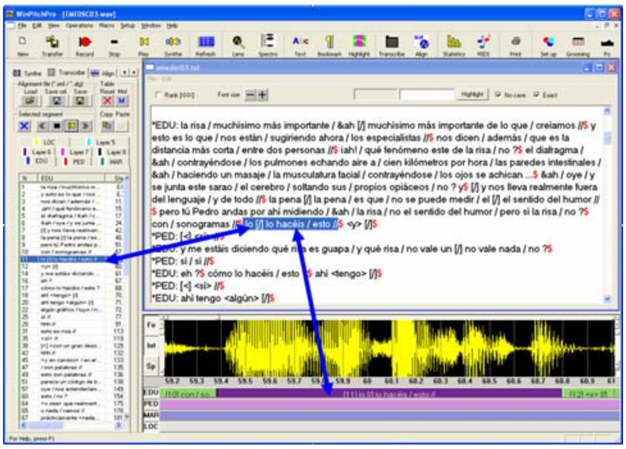
En mil neuf cent soixa...

Figure 1: Simultaneous transcription and alignment. The user sequentially defines segment of speech and enters the corresponding text.

Computer assisted alignment

Experimental studies have shown that coordination between visual spotting of words and positioning of a mouse on a computer screen could be obtain by slowing down speech playback by a suitable factor, depending on the size of the text object to spot (larger chunks of text require less processing time, and thus allow a faster speech playback rate in the process). WinPitch Corpus assisted text to speech alignment is based on this principle.

In the second mode of operation, the text preexists, and is displayed dynamically in a window while the corresponding speech sound is played back at a slower speed (which can be adjusted continuously on the fly). At each identification of a speech unit to segment and align (be a syllable, word, syntagm or sentence), the operator clicks with the computer mouse on the text segment perceived. The program records the position of the cursor on the text window (which defines the end of the text segment to align) and the time of the click (remapped on the real time scale of the speech wave). This process generates continuously a database of pointers linking segments of text and segments of speech.



*EDU: la risa / muchísimo más importante / &ah [/] muchísimo más importante de lo que / creíamos //$ y esto es lo que / nos están / sugiriendo ahora / los especialistas //$ nos dicen / además / que es la distancia más corta / entre dos personas //$ ¡ah! / qué fenómeno este de la risa / no ?$ el diafragma / &ah / contrayéndose / los pulmones echando aire a / cien kilómetros por hora / las paredes intestinales / &ah / haciendo un masaje / la musculatura facial / contrayéndose / los ojos se achican ...$ &ah / oye / y se junta este sarao / el cerebro / soltando sus / propios opiáceos / no ? y$ [/] y nos lleva realmente fuera del lenguaje / y de todo //$ la pena [/] la pena / es que / no se puede medir / el [/] el sentido del humor //$ pero tú Pedro andas por ahí midiendo / &ah / la risa / no el sentido del humor / pero sí la risa / no ?$ con / sonogramas //$ lo [/] lo hacéis / esto //$ <y> [/]$
*PED: [<] <sí> //$
*EDU: y me estáis diciendo qué risa es guapa / y qué risa / no vale un [/] no vale nada / no ?$
*PED: sí / sí //$
*EDU: eh ?$ cómo lo hacéis / esto //$ ahí <tengo> [/]$
*PED: [<] <sí> //$
*EDU: ahí tengo <algún> [/]$

Figure 2: Assisted alignment by slowing down speech playback. At each mouse click on a unit of text perceived at slower speed (top right window), bidirectional pointers are generated automatically between the corresponding speech segment (bottom right window) and a database (left window).

Various tools are provided to backtrack, fine tune speech segment limits (with the help of a displayed spectrogram), dynamically modify limits of overlapping voices, etc.

Slowing down speech

Variable rate speech playback is the core engine of the assisted text to speech aligner. It uses a modified version of the PSOLA algorithm (Moulines et Charpentier, 1990) and allows high quality re-synthesis of natural speech. Its principle is based on the pitch synchronous duplication and insertion of pitch periods in the signal, in order to maintain the original speech spectral characteristics of voiced parts while extending their duration. Unvoiced segments are processed the same way.

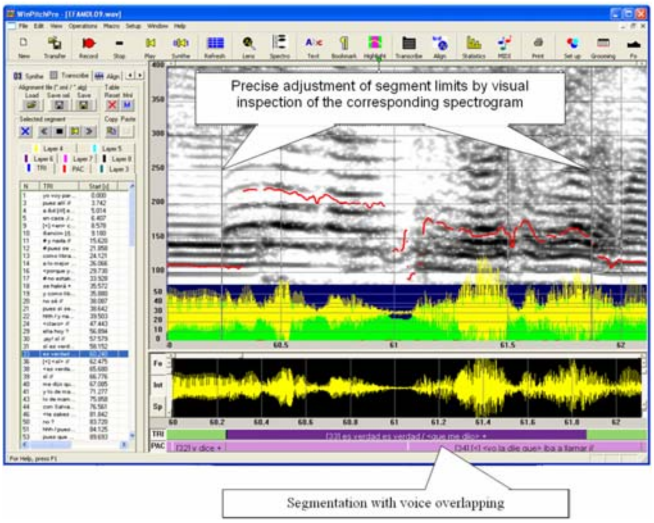
The resulting sound quality is strongly dependent on precise pitch marking, and therefore on reliable fundamental frequency analysis. Errors in pitch marking (missing markers, double markers) induce an echo effect due to the misalignment of pitch chunks added in the PSOLA algorithm. Fo estimation is obtained by the spectral comb method (Martin, 1981, 2000), and the speed playback factor can vary from 7 to ½ (speech played back at double speed). This rate is dynamically adjustable by the user while the alignment is processed, allowing operations on very large files.

Automatic layer assignment

Preexisting text can be organized (following a simple convention for naming speakers turns) so that segments are automatically assigned to their corresponding layers. The user does not have to worry about speaker's turns while aligning, as the program will put segmented text in the appropriate layer assigned to each speaker (8 layers are presently available, but future extension will provide for unlimited number of speakers).

Fine tuning and speaker overlap

Once the assisted alignment session ends, the program displays automatically the text under the corresponding speech segments, represented by their acoustic analysis (spectrogram, intensity and melodic curves, waveform). The user can then adjust precisely the segments by dragging their limits with the mouse, with the help of visual inspection of the corresponding spectrogram (or other available acoustical information). Overlapping segments can then easily and precisely be defined.



Precise adjustment of segment limits by visual inspection of the corresponding spectrogram

Segmentation with voice overlapping

Figure 3: Fine tuning of speech segments limits with the help of a simultaneously displayed spectrogram (which allows precise segmentation in case of speaker's overlapping).

Prosodic morphing

The PSOLA engine is also used for prosodic morphing of any part of the speech data, allowing the user to modify with very simple graphic commands the prosodic parameters or the original data (fundamental frequency, intensity, segment duration, pauses).

Multimodal alignment with WinPitch Corpus

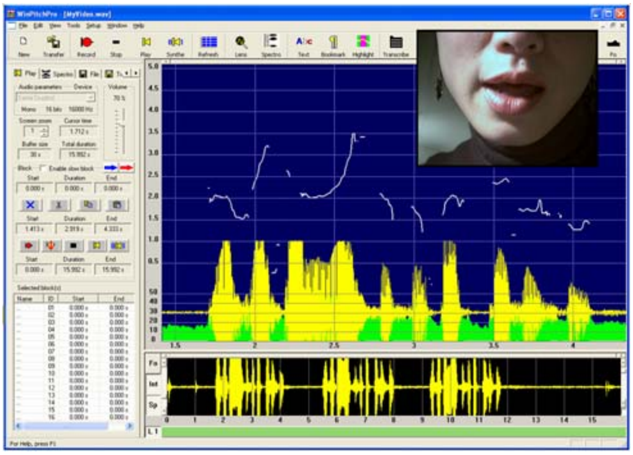


Figure 4: Simultaneous display of acoustic and video information of a multimedia file.

WinPitch Corpus can read and process most multimedia file formats, and offers the same capabilities when displaying video with the sound data. Alignment of multimodal files can be done the same way as with sound files.

The program also allows real time recording and analysis of speech data. This feature is particularly useful to monitor recording from the real time spectrographic display. Problems such as low signal level, echo, saturation can immediately be noticed and fixed.

Conclusion

Computer assisted text to speech alignment, as being much faster than conventional and automatic methods, thanks to the use of ergonomically well designed tools, allows the development of large corpora of various languages, which in turn will, among other benefits, induce a better comprehension of the relationship between syntax and intonation. The method of slowing down speech used in WinPitch Corpus permits to handle corpora of very variable sound quality, which would prevent the use of automatic methods based on speech recognition.

In this software, alignment is executed by an operator clicking on the units to segment (syllables, words, syntagms, sentences) displayed on a program window while the corresponding speech sounds is played back at reduced speed. This slow down process allows the psychometrical coordination needed for the process, which can be executed in one pass et does not require any particular expertise in phonetics. The process is much faster than the traditional approach where a trained phonetician has to align the sound segments one by one by shifting a time window along the speech signal. It is also more reliable that emerging automatic methods, which require reasonably good quality recordings and exclude too large variations or voice and pronunciations characteristics of the speakers.

Numerous functions of WinPitch Corpus allow precise adjustment a segment limits, as well as the exact definition of overlapping segments of speech, thanks to simultaneously displayed acoustic information (spectrogram, waveform, intensity and melodic curves). The user can also edit the transcription on the fly, which any font provided in Unicode. Phonetic transcription, syntactic tagging, and any other information of interest can be easily added on one of the 8 layers available for transcription.

The program has been used extensively in the development of large corpora in four romance languages in the C-ORAL-ROM project (C-Oral-Rom, 2004).

References

Blanche-Benveniste, C. (2002) « Réflexions sur les transcriptions de corpus de français parlé », *Revue PArole*, 22-23-24, 2002, pp. 91-117.

C-Oral-Rom (2004) “Integrated reference corpora for spoken romance languages” (IST-2000-26228 - Shared-cost RTD) http://lablita.dit.unifi.it/coralrom/

Cosi, P. (1997) "SLAM v1.0 for Windows : a Simple PC-Based Tool for Segmentation and Labelling", *Proc. Of ICSPAT-97, Int. Conf. On Signal processing Applications and Technology*, San Diego, CA, Sept. 1997, pp. 1714-1718.

Fohr, D., Mari, J.-F. et Haton, J.-P. (1996) « Utilisation des modèles de Markov pour l’étiquetage automatique et la reconnaissance de BREF80 », *Actes des XXIèmes Journées d'Etude sur la Parole*, Avignon, pp. 339-342.

Malfrère, F. et Dutoit, T. (2000) « Alignement automatique du texte sur la parole et extraction de caractéristiques prosodiques », in *Ressources et évaluation en ingénierie des langues*, Chibout, Mariani, Masson, Néel ed., De Boeck et Larcier, Paris, pp. 541-552.

Martin, J.C. and Kipp, M. (2002) "Annotating and Measuring Multimodal Behaviour – Tycoon Metrics in the Anvil Tool", *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, pp. 29-31 may 2002.*

Martin, Ph. (1981) "Mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne", *Actes des XIIèmes Journées d'Etude sur la Parole*, Montréal, juin 1981.

Martin, Ph. (2000) « Peigne et brosse pour Fo : Mesure de la fréquence fondamentale par alignement de spectres séquentiels », *Actes des XXIIIèmes XXI Journées d'Etude sur la Parole*, Aussois, France, juin 2000, pp. 245-248.

Moulines, E. & Charpentier, M. (1990) "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol 9, pp. 453-467.

Talin, D. and Wightman, C.W. (1994) "The Aligner: Text-to-Speech Alignment using Markov Models and a Pronunciation Dictionary", *Second ESCA/IEEE Workshop on Speech Synthesis*, pp. 89-92.

WinPitch (1996, 2004) http://www.winpitch.com