EASYALIGN SPANISH: AN (SEMI-)AUTOMATIC SEGMENTATION TOOL UNDER PRAAT

Jean-Philippe Goldman, Sandra Schwab
Université de Genève
Jean-Philippe.Goldman@unige.ch, Sandra.Schwab@unige.ch

Abstract

EasyAlign is a user-friendly automatic phonetic alignment tool for continuous speech. It is developed as a plug-in of Praat, and it is freely available. Its main advantage is that one can easily align speech from an orthographic transcription. It requires a few minor manual steps and the result is a multi-level annotation within a TextGrid composed of phonetic, syllabic, lexical and utterance tiers. Evaluation of EasyAlign was performed according to three approaches: a boundary-based, a duration-based and segment-based approach. Results are very promising, showing, on the one hand, little difference between EasyAlign and human alignment, and a good generalization of the training, on the other one. EasyAlign is fully available for French and Spanish, while other languages such as English, Taiwan Min are under development thanks to a growing interest of community users.

Key-words: segmentation, automatic alignment, Spanish, Praat, EasyAlign.

Resumen

EasyAlign es una herramienta sencilla de usar, destinada a la segmentación fonética automática del habla. Es un complemento gratuito de Praat, cuya mayor ventaja es la facilidad con la que segmenta el habla desde una transcripción ortográfica. Tras pocas etapas manuales, el resultado es una anotación en varias tiras que se corresponden con una tira fonética, silábica, léxica y ortográfica dentro un TextGrid. Se ha realizado una evaluación de EasyAlign abordando tres aspectos: las fronteras de los segmentos, la duración de éstos y el propio segmento. Los resultados, muy prometedores, muestran, por un lado, pocas diferencias entre la segmentación automática y la humana y, por otro lado, una buena generalización del entrenamiento del sistema. EasyAlign está disponible para el francés y el castellano; por el momento, para otras lenguas como el inglés y el taiwanés, está en fase de desarrollo, gracias al creciente interés de la comunidad de usuarios.

Palabras clave: segmentación automática del habla, castellano, Praat, EasyAlign.

1. Introduction

The purpose of phonetic alignment (or phonetic segmentation) is to determine the time position of phone, syllable, and/or word boundaries in a speech corpus of any duration, on the basis of the audio recording and its orthographic transcription. Such aligned corpora are widely used in various speech applications including automatic speech recognition, speech synthesis, as well as prosodic and phonetic research.
An accurate segmentation that would be done fully manually would require as many as 800 times real-time; i.e. 13 hours for a one-minute recording (Schiel 2003). The processing time is a major drawback for manual labeling, especially when faced with very large spontaneous speech corpora. Thus, an automatic phonetic alignment tool is highly desirable. Besides, such automatic approach is not only consistent (i.e. has the same precision throughout the corpus)

but also reproducible (i.e. can be repeated quickly and many times). Although an alignment tool can save time, speech, especially spontaneous speech, has many unpredictable phonetic variations that can decrease the accuracy of the process. Even with precise computational tools and data preparation, automatic systems can make errors that a human would not. Thus, post-processing detection of major segmentation errors is needed to improve accuracy. In fact, automatic approaches are never fully automatic nor straightforward and instantaneous as claimed by existing systems. It is a matter of compromise among time, expected precision and computational skills. The question then lies in what is the needed degree of accuracy. For example, corpus-based text-to-speech systems require a high level of alignment precision, but some research studies (at syllable level for instance) may require less precision.

Various computational methods have been developed for automatic phonetic alignment. Some have been borrowed from the automatic speech recognition (ASR) domain. However, the alignment process is much easier than speech recognition because the alignment tool does not need to determine what the segments are but only their location. For this reason, HMM (Hidden Markov Models)-based ASR systems are widely used in a forced-alignment mode for phonetic segmentation purposes (Young et al. 2010). Another approach combines a text-to-speech system (TTS) and a Dynamic Time-Wrapping (DTW) algorithm. In this case, synthetic speech is generated from the orthographic or phonetic transcription and compared to the corpus as in Malfrère (1997). The DTW will find the best temporal mapping between the two utterances using acoustic feature representation. A hybrid system based on these two techniques in cascade (first HMM then TTS+DTW) is presented in Paulo/Oliveira (2004), where results improved. Eventually in van Santen/Sproat (1999), some contour detection techniques borrowed from image processing also give interesting results.

Although these existing systems give good results and are usually freely available, it should be noted that they are not directly usable as the user needs to record speech and train his own acoustic models. Besides, this later preliminary step is usually executed through a command line interface. The presented system, named EasyAlign, relies on HTK (Young et al. 2010), a well-known HMM toolkit. It can be seen as a friendly layer under Praat (Boersma/Weenink 2010), which facilitates the whole alignment process as it comes with a phonetization system and already trained acoustic models.

2. EasyAlign

2.1. GENERAL DESCRIPTION

EasyAlign is a freely available system, made of Praat scripts, but it also includes two external components: a grapheme-to-phoneme conversion system and a segmentation tool for the alignment at the phone level. It is distributed as a self-installable plug-in, and comes with the already trained acoustic models of phones.

The whole process to segment a speech file is as follows: from a speech audio file and its corresponding orthographic transcription in a text file, the user has to go through three automatic steps; manual verifications and adjustments can be done in-between to ensure even better quality. More precisely, these three steps are: macro-segmentation at utterance level; grapheme-to-phoneme conversion, and phone segmentation. The result is a multi-tier TextGrid, the annotation format within Praat, with phones, syllables, words and utterance segmentation as in Figure 1. Providing a TextGrid already segmented into utterances with an orthographic and/or a phonetic transcription speeds up the process as the first macro-segmentation step is possible and can be skipped.

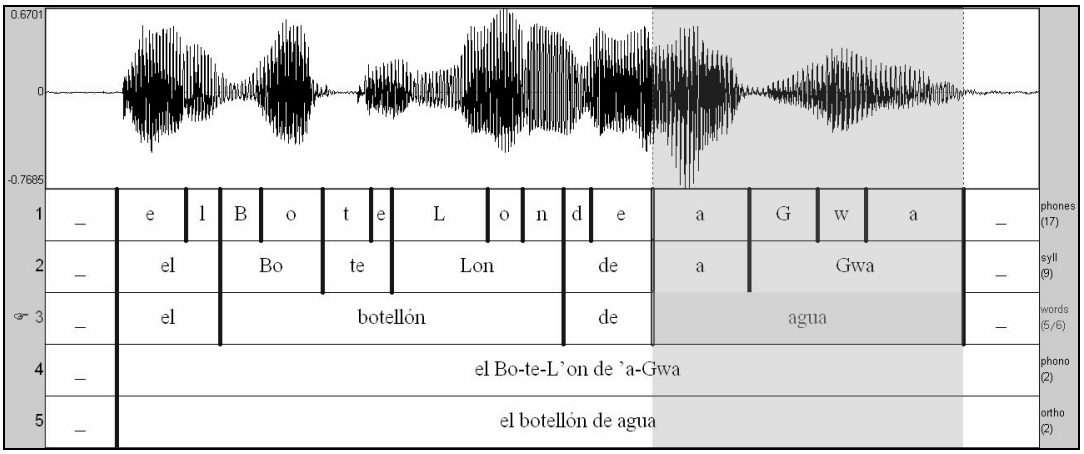


Figure 1. Full resulting TextGrid with 5 tiers from bottom to top: *ortho, phono, words, syllables, phones* for the utterance "el botellón de agua".

2.2. SPANISH EASYALIGN

2.2.1. Development

EasyAlign adaptation to a new language has two needs: some speech data and a grapheme-to-phoneme conversion system. First of all, we recorded, in a sound-treated booth, 6 Spanish native speakers (3 males and 3 females) from the central part of the Iberian Peninsula. Each of them produced between 26 to 30 minutes of reading (total duration = 172 minutes). Then, we integrated the SAGA phonetizer within EasyAlign to convert the orthographic transcription to the phonetic one. This phonetizer was developed at the Centre de Tecnologies i Aplicacions del Llenguatge i la Parla of the Universitat Politècnica de Catalunya (Llisterri/Mariño, 1993; Moreno/ Mariño, 1998). As can be seen in Figure 1 (*phono* tier), SAGA provides a fine phonetic transcription from the orthographic transcription. The phonetic transcription of all productions was manually corrected in order to exactly match the produced utterance (e.g. deletion of /z/ of "las" in the sequence "las ranas", when it was not produced by the speaker). Finally, a stochastic training based on HTK was performed with these speech productions and their corrected phonetic transcription, which resulted in acoustic models for each phoneme. Figure 2 presents the necessary steps to use EasyAlign and to train it.

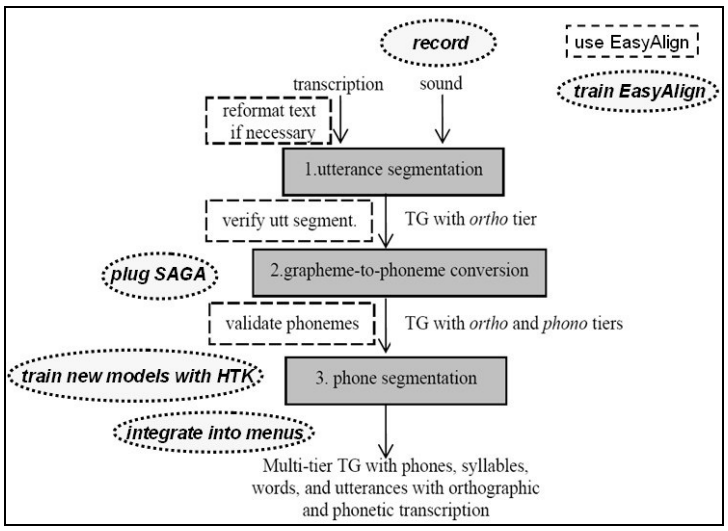
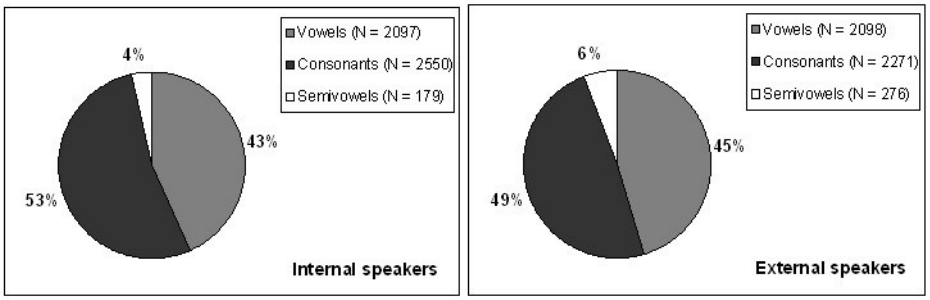


Figure 2. EasyAlign usual process in square boxes as in Goldman (2011) and the adaptation steps in oval shapes.

2.2.2. Evaluation

EasyAlign performance has been evaluated on the basis of a corpus of 12 minutes. One minute of 12 speakers was manually annotated by three phonetic experts (reference alignment) and compared to the automatic alignment. Among the speakers, 6 were "internal" speakers, used in the training corpus and 6 were new "external" speakers, taken from the corpus used in Machuca/Ríos (2008).
Evaluation was performed according to three approaches: a boundary-based, a duration-based and a segment-based approach. In each of the evaluations, only phones were taken into account. Figures 3a and 3b show the repartition of the vowels, consonants and semivowels in internal and external corpus. It has to be noted that in both corpora, semivowels are under-represented, in comparison with vowels and consonants.



Figures 3a and 3b. Vowels, consonants and semivowels repartition in internal and external speakers

Regarding the first evaluation, the boundary-based evaluation, we computed the absolute difference (in ms), for each phone (n = 9471), between the automatic and the manual initial boundaries. Results showed that 60.26% of the differences between automatic and manual boundaries lie within 10ms, and 87.11% within 20ms. Figures 4 and 5 present the distribution of absolute differences between automatic and manual boundaries in internal and external speakers, respectively.

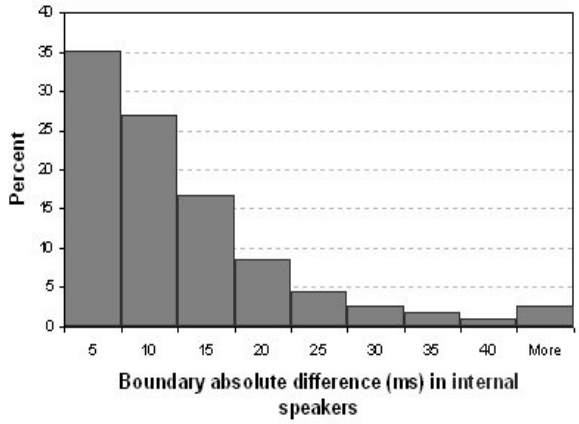


Figure 4. Distribution of absolute differences between automatic and manual boundaries in internal speakers (N = 4826).

Although statistically significant, little difference was observed when comparing the differences within 10ms in internal and external speakers (62.16% and 58.28%, respectively; $\chi^2(1) = 14.92$, $p < .001$). Regarding differences within 20ms, we observe no statistically

significant difference between internal and external speakers (87.65% and 86.54%, respectively; $\chi^2(1) = 2.58$, n.s.).

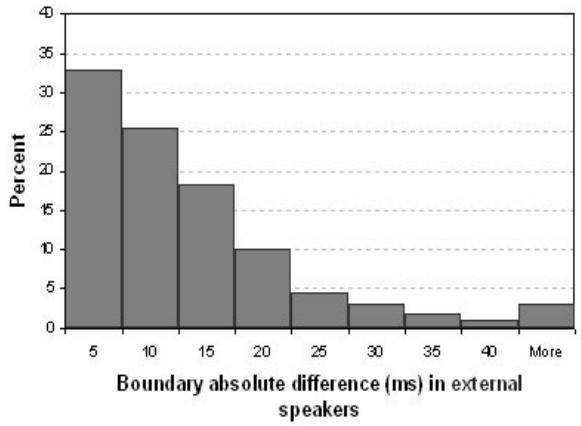


Figure 5. Distribution of absolute differences between automatic and manual boundaries in external speakers (N = 4645).

Then we examined the difference (in ms) between the automatic and manual initial boundaries of the segments as a function of the segment nature (vowels, consonants and semivowels) and as a function of speakers (internal and external). As can be seen in Figure 6, we observe a significant difference between internal and external speakers ($F(1, 9465) = 6.7$, $p < .01$), mainly due to the larger difference found in external speakers for semivowels (5ms) than in internal speakers (1.46ms). Moreover, the difference between automatic and manual segmentation varies according to the segment nature ($F(2, 9465) = 62.12$, $p < .001$). In average, EasyAlign, in comparison with manual segmentation, tends to move vowels boundaries to the right by 7ms, and consonants and semivowels boundaries by 3ms. As can be seen in Figure 6, semivowels present a larger variability than vowels and consonants, whatever the speakers may be. For that reason, no interaction is observed ($F(2, 9465) = 1.72$, n.s.), despite the above mentioned larger difference in semivowels in external speakers than in internal ones.

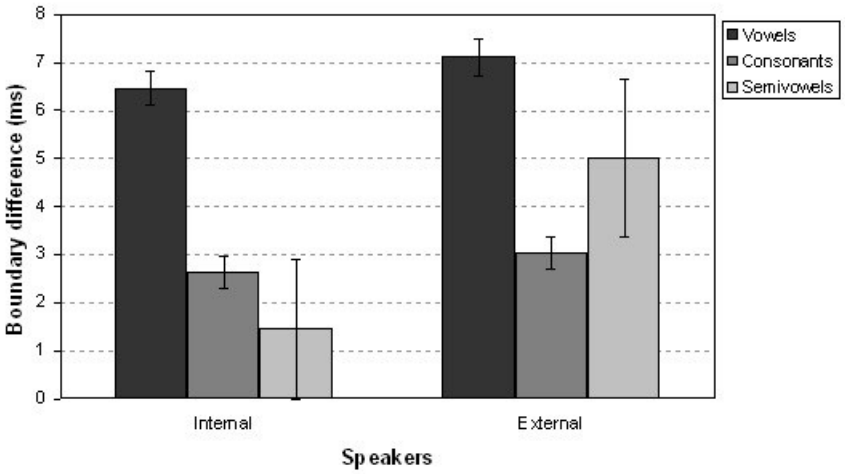


Figure 6. Difference (in ms) between the automatic and manual initial boundaries as a function of segments (vowels, consonants and semivowels) and speakers (internal and external). Error bars are standard errors of the mean.

In the duration-based evaluation, we calculated, for each phone (n = 9471), the difference between the automatic and manual segment durations. Results show a mean difference of

0.11ms (st. dev. = 19.68). We also studied the duration difference between automatic and manual segments as a function of their phonological nature (vowels, consonants and semivowels) and as a function of speakers (internal and external). No significant difference is found between the duration differences in internal (-0.13ms) and external (0.34ms) speakers ($F(1, 9465) = 1.03$, n.s.). Moreover, as can be seen in Figure 7, the duration differences vary according to the phonological nature of the segments ($F(2, 9465) = 159.84$, $p < .001$). In average, EasyAlign, in comparison with manual segmentation, tends to shorten vowels by 4ms and to lengthen consonants by 3ms. Semivowels show the higher duration difference (7.5ms) and a larger variability, which might be explained by their acoustic instability. No interaction is observed despite the larger difference in semivowels in internal speakers than in external ones ($F(2, 9465) = 2.72$, n.s.).

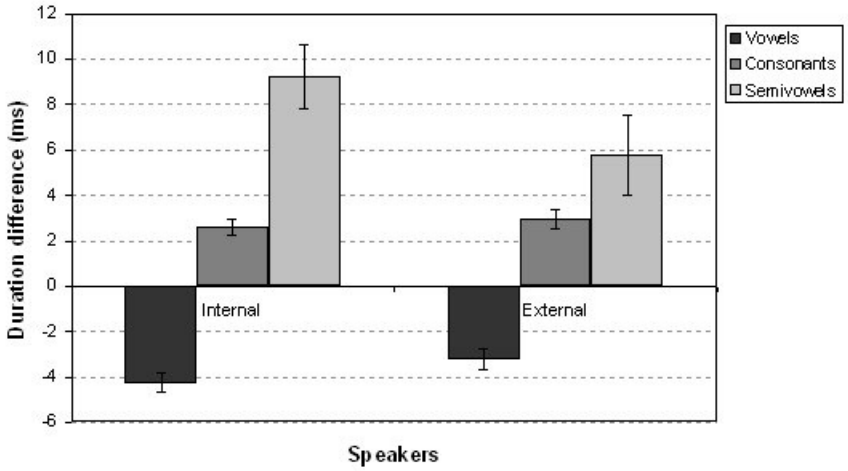


Figure 7. Duration difference (in ms) between automatic and manual alignment as a function of segments (vowels, consonants and semivowels) and speakers (internal and external). Error bars are standard errors of the mean.

Finally, in the segment-based evaluation, we computed, for each phone (n = 9471), the so-called "overlapping-rate", a speech-rate independent measure (Paulo/Oliveira 2004), which represents the ratio between the common part of the automatic and manual segment and the maximal "possible" duration of the segment considering initial and final boundaries of both automatic and manual segmentations. A rate of 0 means that there is no overlap between the automatic and manual segments, while a rate of 1 means that the overlap is total. According to Paulo/Oliveira (2004), a segment with an overlapping rate of 0.75 is considered well segmented.

As the overlapping rate distribution was left-skewed, analyses were performed on medians, with the Mood's Median Test (Mood 1950). Results showed that the median of the overlapping rate reaches 0.73, with little difference, although significant, between internal (0.74) and external (0.71) speakers ($\chi 2(1) = 41.79$, $p < .001$). This difference comes mainly from the overlapping rate of semivowels (see Figures 8 and 9), which is considerably lower in external speakers (0.59) compared with internal speakers (0.64).

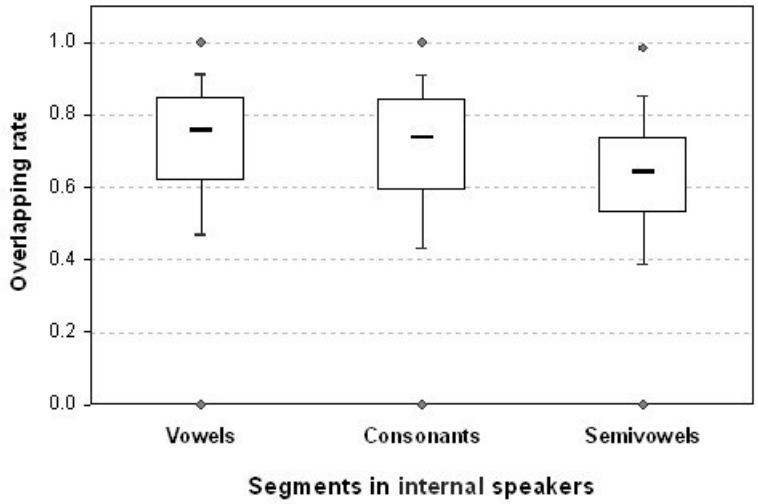


Figure 8. Overlapping rate distribution in internal speakers as a function of segments (vowels, consonants and semivowels).

Let's also mention that semivowels show a lower rate than vowels and consonants, whatever the speakers may be, again by reason of their acoustic instability.

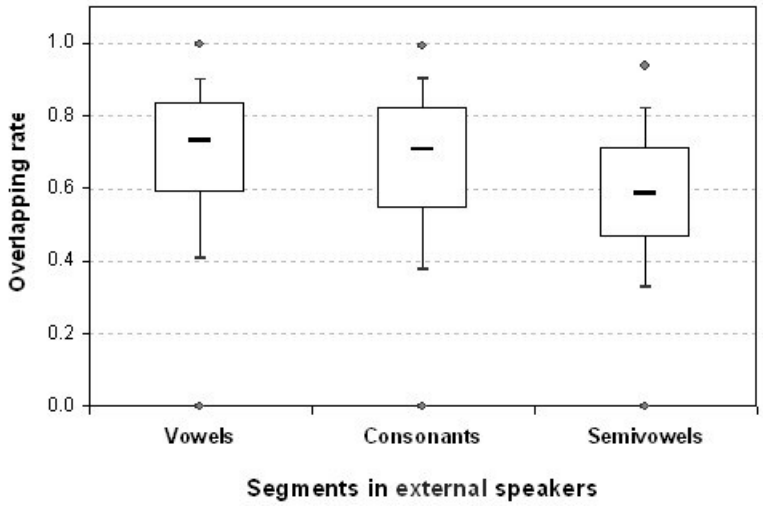


Figure 9. Overlapping rate distribution in internal speakers as a function of segments (vowels, consonants, and semivowels).

3. Conclusion

The three evaluations –boundary-based, duration-based and segment-based– we performed on EasyAlign performance showed very promising results, with little, although sometimes statistically significant, difference between internal and external speakers, which reflects a good generalization of the training.

In addition, results showed that semivowels globally present the greatest difference between automatic and manual segmentation, and a large variability. Two reasons might explain this discrepancy. On the one hand, semivowels, as mentioned earlier, are not easy to segment due to their high acoustic instability and high degree of coarticulation. On the other hand, they were under-represented in the training corpus, in comparison with vowels and consonants. Improvements of EasyAlign performances would be to train it on a corpus with a high

number of semivowels, and to increase the frequency of analysis (which is now of 10ms), as a higher rate of analysis window would yield a higher precision.
In summary, EasyAlign turns out to be an efficient and friendly tool which enables to easily align speech from an orthographic transcription within Praat. To our knowledge, such a tool was not, until now, freely available for Castilian Spanish.

REFERENCES

Boersma, Paul/Weenink, David (2010): Praat: doing phonetics by computer. <http://www.praat.org> (March 2010).

Goldman, Jean-Philippe (2011): "EasyAlign: an automatic phonetic alignment tool under Praat", en: *Proc. INTERSPEECH 2011*, Firenze, Italy (August 2011).

Llisterri, Joaquim/Mariño, José B. (1993): Spanish adaptation of SAMPA and automatic phonetic transcription, en: *Proyecto Esprit 6819. Informe SAM-A/UPC /001/V1* (February 1993).

Machuca, María J./Ríos, Antonio (2008): "Combinaciones de más de tres vocales en los enlaces entre palabras", en: *Actas IV Congreso de Fonética Experimental*, Universidad de Granada (February 2008).

Malfrère, Fabrice/Dutoit, Thierry (1997): "High-Quality Speech Synthesis for Phonetic Speech Segmentation", en: *Proc. Eurospeech 1997*, Rhodes, Greece (September 1997), 2631-2634.

Mood, Alexander M. (1950): *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Moreno, Asunción/Mariño, José B. (1998): "Spanish dialects: phonetic transcription", en: *Proc. ICSLP'98*, Sydney, Australia (November 1998), 189-192.

Paulo, Sérgio/Oliveira, Luís C. (2004): "Automatic phonetic alignment and its confidence Measures", en: *4th International Conference EsTAL* 2004, Alicante, Spain (October 2004).

Schiel, Florian/Draxler, Christoph (2003): *The Production of Speech Corpora*. Munich: Bavarian Archive for Speech Signals.

van Santen, Jan P.H./Sproat, Richard W. (1999): "High accuracy automatic segmentation", en: *Proc. EuroSpeech99*, Budapest, Hungary (September 1999), 2809-2812.

Young, Steve/Kershaw, Dan/Odell, Julian/Ollason, Dave/Valtchev, Valtcho/Woodland, Phil (2010): *The HTK book*. <http://htk.eng.cam.ac.uk> (March 2010).