

Semi-Automatic Aligning of Swedish Forensic Phonetic Phone Speech in Praat using Viterbi Recognition and HMM

Jonas Lindh

Department of Linguistics, Göteborg University, Gothenburg, Sweden

jonas.lindh@ling.gu.se

Automatic alignment of text and sound is of great help and saves a lot of time labelling speech databases, either for research or for developing speech technology tools such as automatic speech recognition or text to speech systems. It is also a very useful tool in forensic speaker identification as one often receives a tapped recording together with an orthographic transcription. The orthographic transcription can be used together with the sound file to provide information of where in the recording significant events occur to a greater or lesser extent. Even if the aligning sometimes is not perfect, it replaces some of the time consuming manual labelling. To perform automatic aligning, common speech recognition techniques are applied at various levels. In this case, a framework for doing automatic aligning, called EasyAlign, was developed for the free software Praat (Goldman, 2007). Praat is distributed as an open source software under a GPL license. On top of the source code a built-in scripting language can execute commands, make calculations and communicate with other programs in different manners (Boersma & Weenink, 2007). To be able to implement a new language for automatic aligning within the framework there is a need for a grapheme to phone converter and a trained Hidden Markov Model that can be used by the viterbi recognition program HVite from the HTK toolkit (Young et al., 2006).

Automatic Aligning of Speech

Aligning recorded speech automatically is a technique that borrows heavily from automatic speech recognition (ASR). Successful attempts have been made using Hidden Markov Models (Brugnara et al., 1993) and dynamic time warping (Malfrère et al., 1998 and 2000), both well-known techniques in ASR. In dynamic time warping, the signal is compared and aligned with a reference from, for example, a text-to-speech system. Using a Hidden Markov Model (HMM) recognition system, forced alignment can be used together with phoneme models and the Viterbi algorithm. (Sjölander, 2001 and 2003). The output of the forced alignment can then be used to create other tiers on other phonological levels. The result can be displayed together with the sound in any software that can read the labelling format, for example Praat.

Hidden Markov Models trained on a Swedish Corpus of Telephone Speech

The Swedish SpeechDat project was part of a larger project to create databases of recorded telephone speech in order to train and develop applications for speech recognition and verification (Elenius et al., 1997). The database consists of telephone recordings made by approximately 5000 people. For a project aimed at developing acoustic models for speech recognition, Giampero Salvi at the KTH trained HMMs on the Swedish SpeechDat database (Salvi, 1999), which he kindly permitted us to use in our project. This HMM is a monophone model (50 monophones) with 8 mixture components.

It uses 39 MFCC:s (including delta and delta deltas). However, this HMM is designed for telephone speech and will not work properly on recordings containing higher frequency components. Therefore, when it is applied to non-telephone recordings, resampling of the recording to 8 kHz is necessary before aligning is applied.

Test Procedure

In forensic speaker identification case work, one often receives a tapped telephone recording together with an orthographic transcription. To save time it would be helpful to get an at least approximate alignment between the text and the speech to find and analyze relevant parts in the orthographic transcription. As test material in the present project, 26 seconds from an authentic recording of a tapped telephone conversation was used together with its orthographic transcription. The transcription was preprocessed by dividing the text into line chunks based on punctuation marks and phrasing. The number of lines determines the number of aligned fragments, i.e. pauses and line breaks should coincide to optimize the alignment.

Results, Conclusion and the Future

An automatic alignment of a sound file and the corresponding orthographic transcription after some minor manual corrections is illustrated in figure 1 below.

Figure 1. Result from an automatic sound-transcription alignment after some minor manual corrections.

The alignment precision at the phoneme level is far from perfect. The vowels are aligned with the highest degree of precision. At this stage in the process, some parts are not aligned for some reason that is not at present fully investigated. At the syllable and word level the aligning is much better and probably the most useful in forensic casework.

We regard the methods developed within the project described here as potentially very useful as a tool to be used in forensic casework, even though there is room for improvement in terms of precision.

References

- Boersma, P. & Weenink, D. (2007) Praat: doing phonetics by computer (Version 4.6.1) [Computer program]. Retrieved May 16, 2007, from <http://www.praat.org>.
- Brugnara, F., Falavigna, D. & Omologo, M. (1993) Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication*, 12, 4, 357-370.
- Elenius, K., & Lindberg, J. (1997). SpeechDat - Speech databases for creation of voice driven teleservices. In Bannert, R., Heldner, M., Sullivan, K., & Wretling, P. (Eds.), *Proceedings of Fonetik 1997*, Dept of Phonetics, Phonum 4 (pp. 61-64). Lövånger/Umeå.