Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *93-96*
Available online at http://www.ling.umu.se/fonetik2003/

An HMM-based system for automatic segmentation and alignment of speech

Kåre Sjölander

Centre for Speech Technology, Dept. of Speech, Music, and Hearing, KTH.

A system for automatic time-aligned phone transcription of spoken Swedish has been developed. Using a speech recording and an orthographic transcription of the words spoken in the recording the system is able to generate a phone-level segmentation without manual intervention. The system uses a technique based on Hidden Markov Models to position 85.5% of all boundary positions within 20 ms of manually segmented reference boundaries on a set of test recordings.

1. Introduction

Transcription of speech recordings at phone-level is a fundamental task in phonetics and speech technology research. Identification of phone segments in a speech material is the starting point for many studies. Typically this is done manually. For concatenative speech synthesis such segmentation provides the foundation for a unit inventory and the accuracy of the segmentation is crucial for the overall quality of the generated synthesised speech. However, it takes skill and considerable effort to produce such phone-level transcriptions. In comparison, orthographic transcriptions can be had at a much lower cost, especially when no time information for word boundaries is needed. Techniques borrowed from automatic speech recognition have been successfully applied to such word-level transcription in order to produce time aligned phone transcriptions automatically (Brugnara, Falavigna & Omologo 1985; Hosom 2001). In this context the method of speech recognition used is usually referred to as forced alignment. The recognition grammar only allows for what actually has been said and the segment boundaries is the output of interest, since the word output will be the same as the input.

This paper describes an automatic speech alignment system developed for Swedish, which is an extension to earlier work in this area (Sjölander 2001). The system takes as input speech sound files and matching text files containing word-level transcriptions of the speech. A specialised speech recognition system, based on Hidden Markov Models, is applied and a time aligned phone-level transcription is produced.

2. System overview

The system has been implemented using a two-level approach. The basic performance critical speech technology algorithms have been put in a function library. In this way they could also be used for other applications with little effort. On top of this library a command line tool has been written, called nAlign. This tool provides a more user-oriented interface. It

has several control options and takes care of the different file formats involved. In most cases it can be used with the default settings for fully automated phone-level time-alignment.

2.1. Library functions

The library functions were implemented as new commands in the Snack Sound Toolkit (Sjölander & Beskow 2000). In this way it was possible to use existing signal processing routines for speech parameterisation and also to benefit from other sound handling functions during the development. The alignment task can be divided into two major parts, parameterisation and decoding.

Parameterisation is the conversion of sound samples into feature vectors. These provide spectral patterns of the speech at equidistant intervals. Typically, from one hundred and up to a thousand vectors are created per second of sound recording.

Decoding is the process of finding the optimal sequence and alignment of a set of Hidden Markov Model phone models and a sequence of feature vectors. Each phone in the utterance is modelled by an HMM. The standard Viterbi algorithm is used with continuous density output distributions. The decoding function takes two inputs, the parameter vectors created when parameterising a recorded utterance and a finite state transition network of HMMs, describing the phone sequence expected in the utterance. In some cases this will be a straight left to right sequence as in figure 1 (top). But in order to handle alternative pronunciations and speech reduction parallel paths in the transition network are used, cf. figure 1 (middle). Similarly, spontaneous pauses can be handled by inserting optional silence models, see figure 1 (bottom). Each state in the network can be assigned one or more text tags. This feature has many applications. It can be used to assign lexical stress-level on vowels or to mark the beginnings of words, paragraphs, or even speaker change in a dialog. Types of information that is not contained in the phone sequence alone can be transferred through the decoding process unaffected and later be used to structure the aligned phones for further processing.

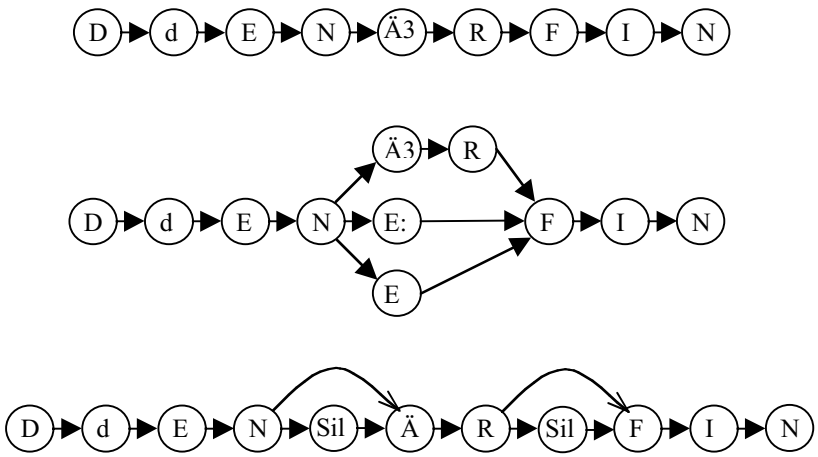

Figure 1. Top graph shows a simple left right transition network. Middle graph shows a network allowing for multiple pronunciations. Bottom graph shows a network that allows for optional inter-word pauses.

2.2. The nAlign tool

The speech alignment tool nAlign, takes sound files and transcription files as input. A transcription file typically consists of an orthographic transcription of the speech contained in the corresponding sound file, but it can also be a phone-level label file if desired. Output is generated as phone label files and, optionally, as word-level files. There is a possibility to use a single transcription file archive, instead of creating individual files. Several options are available to control the tool, for example, what kind of output to generate, if inter-word modelling is to be used, and so on. Also, the tool can be controlled through tags in the transcription files, for example, using the symbols <silence> or <garbage>, to mark pauses or disturbances in the recording. The special tag <fixpoint: *time*> can be used if the alignment process fails somewhere. Using this tag, a certain position in the input text can be locked to the correct time position in the recording manually. This feature can also be used to decrease memory and processing requirements for long recordings. Simple text normalisation functionality has been built directly into the tool. Phone transcription from words is done either through lexicon look-up or through the KTH text-to-speech component RULSYS (Carlson, Granström & Hunnicutt 1982).

3. Results and Discussion

The performance of the system has been evaluated on test material from a manually annotated speech corpus (Bertenstam et al. 1995). The material consists of 327 sentences, with 9318 marked segments at phone-level and a total length of about 19 minutes. The material took 7 minutes to process on a 1.7GHz Pentium IV computer. The system was able to put 85.5% of all boundary locations within 20 ms of the manually segmented reference boundaries. The maximum deviation was 1.214 seconds.

In the development of the automatic alignment tool the goal has been to minimise overall boundary deviation. But in some cases it might be more appropriate to try minimise the maximal deviation. If only a sentence or even paragraph-level segmentation is desired it does not matter much that some phones in the middle of the segment are a few milliseconds off their ideal positions. But the segment boundaries themselves should be reliable. This will be the goal of further research.

The system has been applied to a variety of other tasks, for example, to duration measurements in a study on prosodic boundaries in conversational speech (Heldner & Megyesi 2003) and for measuring speaking rate variation (Bell, Gustafson & Heldner 2003). It should be noted that manual checking and correction was applied after the automatic alignment processing.

4. Conclusions

The automatic alignment system does not perform as well as a trained transcriber for all purposes. But it is fast and convenient to operate and gives consistent results. Naturally, it could be used to create an initial starting point for further manual refinement if desired. The input orthographic transcription needs to reflect the speech as closely as possible in order to give best results. But investigations are underway how to handle speech recording where only part of the speech has been transcribed. The current system applies text normalisation to the input text before processing, for example, every character is converted into lower case or

non-letter characters are removed. In many cases it might be more appropriate to get the original input back but with added time stamps as desired.

The nAlign tool has been used for several purposes and on recordings of varying types with a high degree of success. Even though the resulting alignment is far from perfect the amount of manual labour saved is high. In the future its applicability and performance is expected to continue to grow.

5. Acknowledgements

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. The author would like to thank Mattias Heldner for his many useful comments and suggestions in testing and using the system.

6. References

Bell, L., Gustafson, J. & Heldner, M. (2003) Prosodic adaptation in human–computer interaction. *Proceedings of* ICPhS 2003, Barcelona,

Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., de Serpa-Leitão, A. & Ström, N. (1995) The WAXHOLM application database. *Proceedings of Eurospeech -95*, Madrid 1, 833-836.

Brugnara, F., Falavigna, D. & Omologo, M. (1993) Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication, 12*, 4, 357-370.

Carlson, R., Granström, B., & Hunnicutt, S. (1982) A multi-language text-to-speech module. *Proceedings of ICASSP '82*, Paris, Vol. 3, 1604-1607.

Heldner, M. & Megyesi, B. (2003) Exploring the prosody-syntax interface in conversations. *Proceedings of ICPhS 2003,* Barcelona.

Hosom, J.-P. (2000) Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information. *PhD Thesis*, Oregon Graduate Institute of Science and Technology.

Sjölander, K. (2001) Automatic alignment of phonetic segments. *Working papers 49: Papers from Fonetik 2001*, Lund, Lund University, Dept. of Linguistics, 140-143.

Sjölander, K. & Beskow, J. (2000) WaveSurfer - an open source speech tool. *Proceedings of the ICSLP 2000*, IV, 464-467.