Lund University, Dept. of Linguistics
Working Papers 49 (2001), 140–143

Automatic alignment of phonetic segments

Kåre Sjölander
Centre for Speech Technology, Dept. of Speech, Music, and Hearing, KTH
kare@speech.kth.se

Abstract

Speech data transcribed at the phoneme level is important for basic speech technology applications. This paper describes some experiments with an automatic method for aligning a given sequence of phonemes with the corresponding spoken utterance. It is shown that methods borrowed from the field of automatic speech recognition can successfully be adapted to this problem. Results are reported for experiments that have been carried out on speech data collected in the Waxholm and SWEDIA 2000 projects. In order to confirm the performance level and generality of the proposed method, corresponding experiments have been conducted on the American English TIMIT corpus. For the test utterances from the Waxholm data the best system positions 85.1% of all boundary locations within 20 ms of the manually segmented reference boundaries. For the test utterances from the SWEDIA material, the corresponding figure is 70.9%.

1 Introduction

The availability of speech data transcribed at the phonetic level is crucial in the field of speech technology. For example, when creating a voice for a concatenative speech synthesis system, it is necessary to build a segment inventory from many transcribed utterances from a single speaker. Also, transcribed speech is needed for the training of speech recognition systems and the higher the quality of such transcriptions the better the performance of the resulting speech recognizer. In both cases there is a need for large amounts of correctly segmented and transcribed speech data. Another possible use for large amounts of transcribed speech could be phonetic research. Manual phonetic segmentation of the speech signal is an arduous and time consuming task which imposes a restriction on the amounts of speech that can be produced in this manner. Several methods have been proposed to speed up this process or to make it more or less automatic. The most successful methods have been borrowed from automatic speech recognition, such as Hidden Markov Models (HMM) (Brugnara et al. 1993), or Dynamic Time Warping (DTW) (Malfrère et al. 1998) techniques, because automatic alignment can be viewed as a simplified recognition task. For the HMM case, an HMM recognizer could be used to do forced alignment, that is, a known sequence of phoneme models is used with the Viterbi algorithm to generate a phonetic alignment. In the DTW case, the signal is aligned with some kind of known reference containing the expected segments, for example, generated by a Text-to-Speech System. In both cases it is implied that an orthographical transcription exists, either through manual word level transcription, which is much easier to generate than phonetic transcription, or because the speech was collected by doing recordings of subjects reading text. Transcriptions created by automatic methods also have the attractive properties of being consistent and reproducible. Manual transcriptions still show higher agreement between trained transcribers, than any automatic methods, even though the gap seems to be closing (Hosom 2000).

This paper describes some experiments in using existing speech recognizers for automatic time alignment. Also, some speech parameterization specific modifications are introduced in order to create a recognizer more suited for the alignment task and thus to increase performance. Furthermore, to facilitate a comparison of the proposed method with previous work and in order to confirm the performance level, results are also presented for an implementation based on the American English TIMIT corpus (Garofolo et al. 1990). This corpus has been used extensively in many similar experiments (Hosom 2000).

2 Method and Results

2.1 Test utterances

Two test sets were chosen to evaluate alignment performance, the standard Waxholm test set (Bertenstam et al. 1995) and a number of monologue segments taken from the spontaneous part of the SWEDIA 2000 (Aasa et al. 2000) recordings. The Waxholm test set consists of 327 sentences spoken by 10 speakers, 6 male and 4 female. All utterances were first automatically segmented and then manually corrected and checked. The thirty-four SWEDIA monologues used vary in length, from 14 to 24 seconds, with one speaker per sentence, 17 males and 17 females. They were orthographically transcribed, and then an automatic method was used to create an initial phonemic transcription. This was used as a starting point for manual transcription. 52 phoneme symbols plus one symbol each for silence and noise were used. Both test sets use the same phoneme labels. For stop consonants the closure was assigned a separate label.

2.1 Alignment using existing automatic speech recognition systems

Two existing automatic speech recognition systems were applied to the phoneme alignment task. The recognizer from the ENABL project (Talbot 2000) trained on 16 kHz microphone speech and a SpeechDat (Salvi 1998) recognizer trained on 8 kHz telephone quality speech. Also, for the experiments using the SpeechDat recognizer, that uses a subset of the phoneme labels used for transcribing the test utterances, the transcription files were correspondingly modified. The ENABL recognizer had 69.4% correct boundaries, given a tolerance of 20 ms, as compared to the manual boundaries. Silence-closure boundaries were ignored during the scoring. The same figure for the SWEDIA utterances was 58.9% agreement with manual boundaries. Both test sets were also down-sampled to 8 kHz and tested with the SpeechDat recognizer. The corresponding figures for the test sets were 71.4% and 61.8% respectively for the best acoustic models tried. The results from these tests are also shown in Table 1. The results are lower than for typical state-of-the-art systems, which usually perform between 80 and 90 percent (Hosom 2000).

Table 1. Performance of two automatic speech recognizers on the Waxholm and SWEDIA test sentences. The figures are reported as the percentage of all automatically assigned boundary locations that lie within 20 ms of manually segmented reference boundaries.

Recognizer	Waxholm test utterances	SWEDIA test utterances
ENABL	69.4%	58.9%
SpeechDat	71.4%	61.8%

2.2 Alignment using a modified automatic speech recognition system

A new recognition system was trained using the HTK Hidden Markov Model toolkit, (Young 1997). The training portion of the Waxholm material was used for this. A speech parameterization with Mel-frequency cepstral coefficients, normalized energy, delta, and

acceleration coefficients was chosen. Cepstrum mean normalization was also applied. The frame rate was set to 5 ms, and a Hamming window of size 16 ms was used. This new recognizer gave 85.1% agreement on the Waxholm test set, which compares better with similar systems. For the SWEDIA test set the result was 70.9%. The results of this system are also shown in Table 2.

Table 2. Performance of the new automatic speech recognizer modified for alignment on the Waxholm and SWEDIA test sentences. The figures are reported as the percentage of all automatically assigned boundary locations that lie within 20 ms of manually segmented reference boundaries.

Recognizer	Waxholm test utterances	SWEDIA test utterances
Waxholm aligner	85.1%	70.9%

2.3 Comparative alignment test on American English

The same set-up was used to train an alignment system for American English using the TIMIT corpus. This corpus has been used extensively for similar tasks, (Hosom 2000). The full TIMIT training set was used (3696 utterances) and the full phoneme set was used (61 symbols). The resulting recognizer had 89.9% agreement on the TIMIT core test set (192 files). This compares favorably with the result of 88.9% of the best reported continuous density HMM aligner reported (Brugnara et al. 1993), but not as good as the best system described in the literature (Hosom 2000) at 92.5% agreement, using HMM/Artificial Neural Networks.

3 Conclusions

It is possible to create an automatic phoneme alignment system, through the use of an automatic speech recognizer in forced alignment mode, with some modifications to the speech parameterization. The system described in this paper can generate an alignment that has between 70% and 85% boundary agreement, within 20 ms, with a manual alignment. That the results for the system built using the Waxholm database is worse than the system built using the TIMIT database can probably be explained to a big degree by the fact that the latter database is twice as large. There might also be a difference as to the quality of the transcriptions. It comes as no surprise that the results for the SWEDIA monologues are significantly lower than the results for the Waxholm test sentences. All of the SWEDIA speakers have a pronounced northern Swedish dialect and the speech contains many disfluencies. The Waxholm speakers are mainly from Stockholm and most sentences are short and clear. There is probably room for much improvement. More training data, checked transcriptions, and improved training techniques will be tried. Also, there is much to do in enhancing the alignment process. Currently, a very promising post-processing method is being investigated.

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 organizations.

References

Aasa A., Bruce G., Engstrand O., Eriksson A., Segerup M., Strangert E., Thelander I., & Wretling P., 2000. Collecting dialect data and making use of them: an interim report from Swedia 2000. *Proceedings of Fonetik 2000*, University of Skövde, 17-20.

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*, 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.

Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D, and Dahlgren N. 1990. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology. NTIS Order No. PB91-505065.

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

Malfrère F., Deroo O., & Dutoit T., 1998 Phonetic alignment: Speech synthesis vs. hybrid HMM/ANN. *Proceedings of ICSLP '98*, 1571-1574.

Salvi G., 1998. Developing Acoustic Models for Speech Recognition. *Master of Science Thesis*. TMH, KTH, Stockholm, Sweden.

Talbot N., 2000. Improving the speech recognition for the ENABL project. *TMH-QPSR, KTH*, 1/2000.

Young S., Odell J., Ollason D., Valtchev V., & Woodland P., 1997. The HTK Book. Entropic Cambridge Research Laboratory.