Robust Methods for Automatic Transcription and Alignment of Speech Signals

Leif Grönqvist (lgr@msi.vxu.se)
Course in Speech Recognition

January 2. 2004

Contents

1 Introduction

This is the term paper for the course: “Course in Speech Recognition” given by Mats Blomberg at the Department for Speech, Music and Hearing, KTH (Royal Institute of Technology).

The purpose of this 5 p course is to give students with basic knowledge of speech technology a deeper understanding of techniques for speech recognition.

2 Background

2.1 Göteborg Spoken Language Corpus

For many years, I have been a member of a research group at the Linguistic department at Göteborg University, working with spoken language corpora. Most of our effort has been spent on the so called Göteborg Spoken Language Corpus (GSLC), described in Allwood et al (2000), or Allwood et al (2002) for a paper in Swedish. GSLC is built up to contain many different kinds of social activities rather than different dialects, which is common in spoken language databases. A goal has been to catch how people communicate in natural situations in their everyday life. To get the recording as ecologically valid as possible, [1] the camera and microphones were setup so not to distract the participants.

2.2 MultiTool

Between 1996 and 2000 we had a project aiming to build an environment for a multimodal spoken language corpus (Nivre et al, 1998; Allwood et al, 2001). One main part was to develop a tool for browsing, editing, and alignment of the transcription: MultiTool (Grönqvist, 2000). However, the alignment in the current version of MultiTool is performed manually, which is very time consuming.

Typically a user wants to navigate in a transcription and a media file. Listen to the sound and go to the corresponding utterance in the transcription, and vice versa. The media file may be up to a few hours with a transcription containing maybe 30 000 running words. Therefore, it is very important to be able to navigate in the files in an easy way. To do this, syncronization points at every utterance start, and also on intervals of for example ten seconds are necessary if there are long utterances. A word by word alignment is overkill here.

This paper describes how automatic alignment could be added to MultiTool as a way to simplify the tedious manual work to make the alignment.

3 Automatic Alignment and Annotation

Basically, the needs in a tool like MultiTool leads to two possible settings for automatic alignment depending on the task:

- A media file + transcription as input and alignments as a result (automatic alignment)

[1] Actually, if you want to record a conversation at for example a travel agency, they will absolutely not let you attach microphones to the customers

- Media file as input and a transcription aligned to the audio file as output (automatic transcription and alignment)

Automatic transcription seems to be much more difficult, simply because it has less information to work with. On the other hand, the automatic alignment task could be even more difficult if the transcription does not match the media signal, and the aligner is forced to accept the input transcription.

Actually, both these tasks are far to difficult with the kind of recording we have in the GSLC corpus. Because of the need for ecological valid recordings, both the audio and the video is in quite bad quality compared to what normal ASR software need to work properly. Some important properties are:

- Long distance between speaker and microphone
- Many speakers in one speech signal
- The speech is spontaneous, it contains:
  - Disfluencies
  - Repairs, repetitions, deletions
  - Fragmental speech
- Simultaneous speech is common
- A large number of participating speakers, 1-20 per recording with an average of 4.8 with a median of 3

All these properties lead to a challenging audio signal which is almost impossible to handle for an ordinary open vocabulary ASR (Automatic Speech Recognition) dictation program. But even partial speech recognition could be very useful:

- Find the time points in the speech signal when utterances start or end
- Guess the speaker of each utterance from a list of speakers with known voice properties (obtained by training on an audio file containing many different speakers)
- Find the time points for some easy to recognize sounds or words
- Find the time points for silent or non-speech sections
- Find out if two utterances are uttered by the same speaker
- Find the start- and end time for a given transcribed segment

We will not address all these problems directly in this article but they are related, so models improving one of them could be helpful for the others as well.

4 Approaches to Automatic Alignment and Annotation

According to Schiel et al (2003), only a few fully automatic methods have given usable results so far. These are:

- Segmentation into words: if the words are known and the speech is not very spontaneous
- Markup of prosodic events (Tobi annotations)
- Time alignment of phonemes using HMM:s
- Segmentation and labelling into phonetic units using HMM:s for the words and statistical pronunciation rules
- The "elitist approach" developed by Steve Greenberg. Yields a stream of articulatory features

Let us take a look at means to achieve these goals.

4.1 Sentence Boundary Tagging

Stolcke & Shriberg (1996) report results from an experimental tagger of sentence boundaries (including utterance boundaries). The method is to calculate probabilities for boundaries and non-boundaries between each word and then the Viterbi algorithm is used to find the most probable sequence of boundaries/non-boundaries based on word n-grams. The results were improved in a second experiment using part-of-speech tags. We should note that their experiments were performed using the Switchboard part of the Penn Tree-bank, which has a very good sound quality and separated speech signals for the speakers, and that it was trained using supervised learning.

4.2 Inter-word Event Tagging

For our purposes, Stolcke et al (1998) is more interesting. The authors of this paper have tried to define a tagger for inter-word events: Sentence boundaries, filled pauses (disfluencies), repetitions, deletions, repairs, and ordinary fluencies (when none of the other phenomena appear) between words. As an input they used: the transcribed words, time marks from a forced alignment from the SRI Decipher speech recognizer, and some acoustic measures: fundamental frequency (f_0), phone duration, and signal-to-noise ratio values. In the training phase, a corpus hand-annotated with disfluencies and sentence segmentation was used. Besides the usual acoustic models and language models used for ASR, they used statistical models for the prosodic features, and the inter-word events. This tagger has been tested on the Switchboard corpus, and the approach to combine the various information sources gives an improvement compared to individual sources.

However, both these tagging tasks have been tested using high quality speech signals on separated channels, without overlapping speech.

4.3 HMM-based Segmentation and Alignment

So called forced alignment finds the most probable alignment for a sequence of words given an audio signal. Each word is aligned to a corresponding interval in the speech signal. It could also be used to align down to phoneme level.

In Sjölander (2003), a system for fully automatic alignment is described. This seems to be exactly what we need! The system takes sound files and corresponding text files containing word-level transcriptions of the speech. As an output we get aligned phone-level transcriptions. The intended use seems to be to build up databases useful for many tasks is speech technology, i.e. speech synthesis. Unfortunately we have not seen any results from experiments using low quality audio files.

5 Features Useful for Robust Alignment

Most people (at least all we are aware of!) working with automatic segmentation and alignment use audio files of good quality. There are at least two good reasons for this:

- They need the alignment for speech synthesis or recognition and therefore the speech is non-spontaneous and recorded in a studio

- Otherwise the methods would work very poorly – after all, they are designed for high quality audio input

However, we are interested in aligning lower quality audio as well, so the design of the system has to be changed in some way. Let us first take a look at some useful features we could make use of (Hossom, 2000):76.

5.1 Intensity Discrimination

A nice thing with intensity is that it is easy to measure, but changes in intensity may still be useful as indicators to phoneme-changes, word boundaries, etc. They may also be used for voicing determination, glottalization and impulse detection.

5.2 Voicing Determination and Fundamental Frequency

The most obvious way to discriminate between voiced and unvoiced sections of the speech signal is to use the Cepstrum. If we have a voiced signal, we expect the energy in specific frequencies to remain over time to a much higher extent than for unvoiced signals, which will be shown by the Cepstrum. Other more complicated methods exist, like combining many features weighted together.

Many of the methods for voicing determination also give the fundamental frequency (f_0), and if we have an f_0 we also know that the signal is voiced.

5.3 Glottalization and Impulse Detection

Automatic methods for glottalization and impulse detection are not used very much but should be possible to build. The intensity in the signal together with the fact that f_0 decreases fast before glottalization (Hossom, 2000):94 could be used for detection. For impulse detection we may look at sudden increases in intensity.

6 Robust Alignment

We think that the existing methods to detect features like f_0 and glottalization are useful, besides ordinary phoneme probabilities, but it is important to be able to give a probability how sure we are for each occurrence of the feature. The error rate will of course be very high with this kind of input, but if we are able to pass on the difficult choices, a high precision on the cost of a lower recall may be obtained. This is just fine, because we do not need to align them all anyway.

6.1 A problem

One overall problem is that we want to be able to handle a very noisy signal. The noise may in some cases be simultaneous speech. An interesting approach to handle this would be to reformulate the probabilities. Instead of first trying to filter away the noise and then calculating probabilities that this feature vector corresponds to a specific phoneme, we could train the model on noisy signal to give a probability for the phonemes, given the noisy input.

6.2 A naive attempt

One drawback with the approach to use features well known to work on high quality audio now when we want it to work on low quality audio! So, we would like to revise the ordinary HMM speech recognition.

Instead of phonemes in the states and a posteriori probabilities for the phonemes depending on feature vectors, we would like to try to use ordinary letters in the states. Then we can use the Viterbi algorithm to find the most probable sequence of letters (space included), including alignment, which is exactly what we want.

With partly unsupervised training [2] we could train probabilities for letters depending on feature vectors. Some letters does not really correspond to an interval in the sound signal, but we could expect to find good distributions for most of the letters. The advantage with this kind of training is that we could find dependencies between letters and frequency features that we could not guess in advance, but also that we do not have to rely on phonetic transcriptions for training. If the use of letters rather than phonemes seems to give a much worse result than expected for some specific sounds, specific symbols representing more than one letter may be added.

[2] We have the correct sequence of symbols but not the duration in each state

7 Conclusion

We have proposed some ideas on how to find probability distributions useful for robust alignment and recognition

Unfortunately we have not been able to run any experiments due to lack of corpus data, preprocessor software, a good microphone, etc. Probably the Snack Package may be useful together with the Hidden Markov Model Toolkit (HTK), but without any experience on these packages, it would take too much time to do the experiments.

One problem mentioned earlier that we have not adressed a soluton for, is the ability to find when an unknown speaker starts an utterance. Features used in ordinary speaker recognition should be useful here as well, but the cases with simultaneous speech must be handled in a good way.

References

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. & Ottesjö, C. (2000), 'The Spoken Language Corpus at the Linguistics Department, Göteborg University', *Forum Qualitative Social Research*.

Allwood, J., Grönqvist, L., Ahlsén, E. & Gunnarsson, M. (2001), Annotations and Tools for an Activity Based Spoken Language Corpus, *in* '2nd SIGdial Workshop on Discourse and Dialogue Proceedings, Aalborg, Denmark'.

Allwood, J., Grönqvist, L., Gunnarsson, M. & Ahlsén, E. (2002), 'Göteborgskorpusen för talspråk', *Nydanske Sprogstudier, NyS 30 (special issue on "Korpuslingvistik"),* **NyS 30**, 39–58.

Grönqvist, L. (2000), *The MultiTool User's Manual*, Göteborg University, Department of Linguistics.

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

Nivre, J., Allwood, J., Holm, J., Lopez-Kästen, D., Tullgren, K., Ahlsén, E., Grönqvist, L. & Sofkova, S. (1998), Towards Multimodal Spoken Language Corpora: TransTool and SyncTool, *in* 'Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Speech at COLING-ACL '98'.

Schiel, F., Draxler, C., Baumann, A., Elbogen, T. & Steffen, A. (2003), *The Production of Speech Corpora*, Web Publication.

Sjölander, K. (2003), An HMM-based system for automatic segmentation and alignment of speech, *in* 'Proceedings of Fonetik 2003', pp. 93–96.

Stolcke, A. & Shriberg, E. (1996), Automatic Linguistic Segmentation of Conversational Speech, *in* 'Proc. ICSLP '96', Vol. 2, Philadelphia, PA, pp. 1005–1008.

Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tur, G. & Lu, Y. (1998), 'Automatic detection of sentence boundaries and disfluencies based on recognized words'.