AUTOMATIC LABELLING OF SPEECH GIVEN ITS TEXT REPRESENTATION

Mats Blomberg and Rolf Carlson
Department of Speech Communication and Music Acoustics, KTH
Box 70014, S-10044 Stockholm

A system for phonetic and word level labelling of speech signals is being developed. It features completely automatic generation from the text representation of the utterance to the phonetic transcription and the time position of the phonetic units. This reduces the amount of manual work required for labelling of large speech corpora. Optional pronunciation is predicted and stored as a pronunciation network. The optimal transcription is chosen during the acoustic match to the utterance. A speech corpus consisting of 2000 sentences spoken by one male speaker is currently being labelled at the department using this system. Errors in the resulting phonetic string and in the time positions of the labels are manually corrected. Initial results are presented.

INTRODUCTION

Large, labelled speech corpora are necessary for the training of speaker-independent large-vocabulary recognition systems as well as for other types of phonetic research. The labelling of these corpora requires much manual work, even if semi-automatic procedures are used for aligning the utterances to their corresponding phonetic transcriptions. These procedures require a correct phonetic transcription, obtained through careful listening and inspection of spectrograms by phonetic experts.

The tedious process of manual preparation of the transcriptions can be replaced by techniques for automatic generation of phoneme transcriptions from text. However, there are often several "correct" pronunciation alternatives of a given text and it is impossible to predict the choice of the particular speaker. Therefore, all these alternatives should be considered by the acoustic match during the labelling procedure.

Another problem is that existing rules for grapheme-to-phoneme conversion do not produce a completely error-free transcription. However, if the rule designer is uncertain whether to apply a specific rule in a certain context, he may make it optional, i.e., there will be two resulting strings after applying this rule. The choice between the two pronunciations can be performed in the acoustic match during the alignment procedure.

We have previously developed techniques for automatic speech labelling (Blomberg, 1990). In this report, we have extended that system, taking into account the above considerations.

A system for speech labelling incorporating these techniques can be regarded as a recognition system. The vocabulary is, of course, quite small and the language is extremely limited, but at the acoustic-phonetic level there may be quite difficult phonetic decisions to be made. Also, the demands on a labelling system is higher regarding the required precision in the time positions of the segment boundaries. The output of recognition systems is typically a sequence of words and it doesn't really matter to the user if a correct phonetic transcription of the word is used or if the phoneme boundaries are correctly placed. However, the word recognition performance is highly dependent on the phoneme recognition accuracy. Therefore, it is worthwhile to work on improving techniques of phonetic labelling, even from the point of speech recognition development.

METHOD

Multi-level labelling

We have chosen a hierarchical sentence representation structure with tying between the elements of the different levels. Therefore, an alignment to the phonetic transcription automatically determines the boundaries of the individual phonemes and words of the sentence. The produced label file contains information on the time position of the words of the sentence as well as the phonetic elements.

Generation of transcription network

A rule set designed for text-to-speech applications (Carlson et al., 1982) is used for the generation of a baseform phoneme transcription of an utterance. Optional pronunciation forms are then created through lexicon lookup at the sentence, word, phoneme, and the acoustic-phonetic levels. They are added to the baseform transcription string, which then is turned into a pronunciation network. In a future version, the lexicon lookup procedure will be replaced by rule formulation.

The optional pronunciation choices of the sentence can be based on various types of information. If inserted silent intervals in an utterance are not foreseen, severe alignment errors may occur. The phonemes at either side of the pause have to be connected and therefore will contain the silent interval. Quite simple rules for optional silence at phrase or word boundaries can improve the labelling accuracy significantly. This is especially valid for longer sentences, in which the speaker needs to inhale in the middle of the utterance. At the word level, numbers and abbreviations may have several possible realisations. Sometimes the best form is predictable from the context, but often

more than one pronunciation form is allowed. Other elements of high variability are function words that often get severely reduced.

Acoustic-phonetic representation

The prototype units used by the alignment procedure are context-dependent phones, represented at a production parameter level. The parameter movements within the phone are approximated by piece-wise linear segments. The spectral description, which is used during the alignment, is generated through a speech production mechanism. An argument for using the production parameter representation is that the voice source characterisation is separate from the vocal tract filter. This enables transformation to new speakers and allows dynamic adaptation to the speaker's voice, thereby improving the labelling accuracy.

Phonological rules

Pronunciation alternatives can be predicted by a rule system, similar to the one used by Carlson et al. (1982). The main difference is that the rules are optional and therefore generate a transcription network instead of a unique string. For optimum performance, probabilities should also be assigned to each rule. This module is not yet implemented in the system.

Special rules may be required for function words and common word endings. To account for this fact, a rule system must have access to the word identity and maybe also the sentence needs to be parsed into parts-of-speech. At this stage in the system development, we have a small lexicon with pronunciation alternatives of the most common words.

Training

The context-dependent phones are trained using an analysis-by-synthesis technique for extraction of synthesis parameters. We have used 200 sentences for training, recorded by the same speaker as used for reading the test corpus. In this material, 59 phones, 1979 diphones and 3295 triphone units have been found. Each phone unit is divided into a number of spectral subsegments. Their spectral statistic distributions are generated through a synthesis procedure from the statistical distribution of the synthesis parameters.

Alignment procedure

Before the start of alignment of an utterance, each phoneme in the generated transcription is assigned to a phone unit in the reference material. The best choice is the corresponding triphone. However, it may not have been observed in sufficient number in the training material. In that case a diphone unit is picked if that has occurred often enough. Otherwise the context-free phone is picked. During the alignment procedure, Viterbi search is performed in the spectral domain. The duration distribution of the substates are used explicitly.

EXPERIMENT

Automatic labelling is currently being performed on a set of 2000 sentences recorded by one male speaker. Phonetically trained persons at the department correct the automatically produced labels and their positions. The corrected label files are then compared to the original files

PRELIMINARY RESULTS AND DISCUSSION

The system is being gradually developed and improved during the task of labelling the described corpus. Therefore, detailed results cannot be presented here. A small evaluation on 19 sentences processed by the latest version of the system has been made, though. The sentence length varied from 3 to 29 words. The average magnitude misalignment was 15 ms. 90 % of the labels are within 35 ms from the correct position. This is better than what was reported by Blomberg (1990), even though the phonetic transcriptions of that speech material were manually corrected. Therefore, these preliminary results suggest that automatic text-to-phoneme conversion can give equal labelling performance as manually corrected transcriptions.

ACKNOWLEDGEMENT

This work has been produced in the Swedish Language technology programme, financed by NUTEK and HSFR

REFERENCES

Blomberg, M. (1990): "Automatic detection of the phoneme boundaries in an utterance given its phonetic transcription", *PHONUM - 1, Papers from Fonetik-90*, Umeå.

Carlson, R., Granström, B. and Hunnicut, S.(1982): "A Multilingual Text-to-Speech Module", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Paris.