24 Speech Recognition for the Hearing-Impaired

Panu Somervuo

Presently there are communication services available in Finland based on human speech-to-text (STT) and text-to-speech (TTS) interpretations, e.g. in telephone network and similar services in meetings attended by deaf persons. Such aid, based on human interpreters, is expensive and often problematic due to the intimate discussions that are interpreted.

The progress in speech technology has opened possibilities to automate these tasks. Finnish language appears to be very well suited to automatic STT conversion because the mapping from phonemes into graphemes is straightforward and the phonemic speech recognition has given promising results [2]. Therefore it could be possible to construct communication aids for the deaf and hard-of-hearing persons by using computer-based speech-to-text (STT) and text-to-speech (TTS) conversions.

By using a good phonemic speech recognizer in the STT conversion, the final word and content recognition could be left to the subject reading the raw output of a speech recognizer (grapheme string) on a screen. The other conversion direction, i.e., TTS synthesis, is no technical problem; several synthesizers exist for Finnish.

Recognition score requirements for STT conversion were assessed in the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology by simulating the reading of recognized messages. It was found that for isolated words the comprehension is good up to a 10 % phoneme error rate, for sentences up to 10-20 %, and for dialog sentences up to even 25 %. These results defined requirements for the recognition rate in the present application domain.

24.1 Experiments

The speech database was collected from 12 male speakers and 5 female speakers. The baseline speech recognition experiment was done with speech having 16 kHz sampling rate. A speaker-dependent speech recognizer [1] based on semi-continuous hidden Markov models was trained separately to each speaker. A 30-dimensional feature vector consisted of three concatenated mel-cepstra and the time window for its computation was 100 ms. Three first speech sets of each speaker were used in the training of the system and the fourth speech set was used for testing it. One speech set consisted of 350 Finnish words. Recognition error was computed as a number of inserted, deleted and changed phonemes in the recognized word divided by the number of the phonemes in the correct word spelling. The average phoneme recognition error was 9.1 % for male speakers and 8.9 % for female speakers. Another measure was computed as an amount of correct phonemes in the recognized word. These numbers were 93.2 % and 93.9 % for male and female speakers, respectively. Examination of the speech database revealed that there were missing endings and not well articulated words in some speakers' speech so that some of the phoneme errors using this speech database were due to the errors in the data.

For the evaluation of the speech recognition using analog telephone, the content of the database was filtered to the frequency range 300 Hz - 3400 Hz and downsampled to 8 kHz. The average phoneme errors were now 10.6 % for male speakers and

11.0 % for female speakers. The amount of correct phonemes was 92.0 %.
Some attempts towards the speaker-independent speech recognition were also made because this is a highly desirable feature in the target applications. As a by-product of this, new speaker clustering method was proposed. When the Self-Organizing Map is used as a codebook of each phoneme for each speaker, the similarity between two speakers can be defined as a distance between the phonemewise codebooks of the speakers. This allows the mapping of speakers into two-dimensional plane so that similar speakers are located near each other and speaker clusters can then be easily visualized.

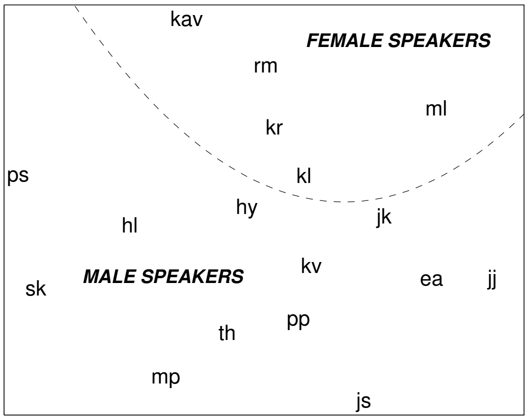

Figure 40: Speakers denoted by their initials are mapped into a two-dimensional plane so that similar speakers are located near each other. The similarity measure is based on the phonemewise Self-Organizing Maps of the speakers which form the basis of the speech recognition. The manually drawn dashed line shows that male and female speakers are discriminated.

The experiments reported here were done in the Neural Networks Research Centre as a feasibility study belonging to the project of the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology. The objective of this project was to experimentally investigate how different phonemic recognition schemes could be used in speech-to-text conversion aids for the hearing-impaired. Other set of experiments was done in the Speech and Audio Systems Laboratory, Nokia Research Centre. To our knowledge this was the first study where phonemic recognition was evaluated and shown to be a potential method for practical speech-to-text aids of the hearing-impaired.

References

[1] Kurimo, M. (1993) Using LVQ to enhance semi-continuous hidden Markov models for phonemes. Proc. of 3rd European Conf. on Speech Comm. and Tech., vol 3. pp. 1731-1734 20.

[2] Torkkola, K., Kohonen, T. et al. (1991) Status report of the Finnish phonetic typewriter project. In Kohonen, T. et al., editors, Artificial Neural Networks, volume 1, pp. 771-776. North-Holland.