APPLICATIONS FOR THE HEARING-IMPAIRED: EVALUATION OF FINNISH PHONEME RECOGNITION METHODS

Matti Karjalainen†, Peter Boda††, Panu Somervuo†††, and Toomas Altosaar†

†Acoustics Laboratory
Helsinki University of Technology
P.O.Box 3000, FIN-02015 HUT, Finland
E-mail: matti.karjalainen@hut.fi

††Speech and Audio Systems Laboratory
Nokia Research Center
P.O. Box 100, FIN-33721 Tampere, Finland
E-mail: peter.boda@research.nokia.com

†††Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 2200, FIN-02015 HUT, Finland
E-mail: panu.somervuo@hut.fi

ABSTRACT

It has been hypothesized that the Finnish language is well suited to speech-to-text conversion for the communication aids of the hearing impaired. In a related study it was shown that, depending on context, 10 to 20 % of phoneme errors can be tolerated with good comprehension when reading text converted from raw phonemic recognition. Two sets of phoneme recognition experiments were carried out in this study in order to evaluate the performance of existing speech recognition systems in this application. For telephone bandwidth speech both systems showed speaker dependent error scores of about 10 % or below, thus supporting the feasibility of the application. For speaker independent cases the error rate was typically more than 20 % which is too high for effortless and fluent communication.

1. INTRODUCTION

The application idea motivating this study is based on the relatively high recognition score of phonemic speech recognition and the almost one-to-one correspondence between spoken and written forms in the Finnish language. This suggests that it could be possible to construct communication aids for the hearing impaired, based on speech-to-text (STT) and text-to-speech (TTS) conversions.

Three kinds of communication aids have been considered based on STT and TTS techniques. First, an automatic STT conversion for a deaf person and TTS conversion back to a normal hearing person over a telephone line could be implemented. As will be shown below, for a limited set of pre-trained speakers this is already feasible now. As another application, an automatic STT interpreter for deaf subjects could be applied for meetings attended by normal and deaf persons. The limited speaking ability of deaf subjects as well as external noise may make the application difficult. As an ultimate form of STT conversion, a portable personal device could do the speech-to-text interpretation.

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

The Finnish language appears to be very well suited to automatic STT conversion on the phonemic level for two reasons. First, a string of phonemes is very easily mapped to text so that only in very rare cases do problems appear. Second, it has been shown that phoneme recognition scores of up to about 95 % are feasible with existing methods [9]. Thus, the STT conversion could be done with a good phonemic speech recognizer, leaving the final word and content recognition to the subject reading the raw output (grapheme string) on a screen. The other conversion direction, i.e., TTS synthesis, is no technical problem; several synthesizers exist for Finnish.

In another part of the study we have assessed the recognition score requirements by simulating the reading of recognized messages with phonemic errors [1]. Random deletions, insertions, and replacements of phonemes, both in-class and intra-class, were generated into isolated words, isolated sentences and dialog texts. The comprehension and the reaction time of reading such erroneous messages as a function of phonemic error rate were tested with a set of subjects using a computer program. As a result we found that for isolated words comprehension is good up to about a 10 % error rate, for sentences up to 10-15 % errors, and for dialog sentences up to about 20 %. This result sets bounds for the recognition of STT conversion in the present application domain.

2. EXPERIMENTS AT NNRC

2.1. Speech corpus and baseline experiments

Currently there exists no public speech database for Finnish. The experiments in this evaluation were done with speech collected from 12 male speakers and 5 female speakers. Each speaker had uttered 350 Finnish words on four different days. The speech had been collected with a 16 kHz sampling rate and quantized to 16 bits per sample.

The first set of experiments including full bandwidth recognition, telephone bandwidth recognition, and speaker clustering was done in the Neural Networks Research Centre at the Helsinki University of Technology. The speech recognizer was similar to that described in [4], which was a further modification of the speech recognizer described in [9] using semi-continuous HM models. Each phoneme was modeled by a five-state left-to-right model using a 7×10-unit Self-Organizing Map [3] as a basis of the state probability density function (pdf). A spherical Gaussian kernel was attached to each SOM unit and a common fixed kernel width was used as a smoothing parameter of the pdf. HMMs were trained by the segmental K-means algorithm [6] and no fine tuning was done by LVQ in these experiments. During the training only state transition probabilities inside phoneme models were re-estimated. The probabilities for phoneme transitions had been estimated from a larger Finnish text corpus.

The baseline experiment was performed by training a speaker-dependent speech recognizer separately for all speakers. Three speech sets of each speaker were used in the training of the system and the fourth speech set was used for testing it. In earlier studies [9, 4, 8] as many as 20 mel-scale cepstral coefficients had been computed from a single speech frame for a feature vector but now this number was halved. The feature vector used in the experiments consisted of three concatenated mel-cepstra. The concatenation was done with 10-dimensional mel-cepstra 50 ms apart from each other so that the time window for computing one final feature vector was 100 ms [8].

Recognition error was computed as the number of inserted, deleted and substituted phonemes in the recognized word di-

vided by the number of the phonemes in the correct word spelling. The average phoneme recognition error was 9.1 % for 12 male speakers and 8.9 % for 5 female speakers. Another measure was computed as the number of correct phonemes in the recognized word divided by the number of phonemes in the correct word spelling. The average correctness measures were 93.2 % for male speakers and 93.9 % for female speakers. It may be useful to observe both of these percentages. If the error percentage is high but also the number of correct phonemes is high it means that there are additional inserted phonemes in the recognized words, but also that the correct phonemes are present. 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 training and testing data. Some errors were also due to the misclassifications between single and double phonemes. Simple duration threshold was used for each phoneme in determining whether to use the short or longer version of the phoneme in the recognized word.

2.2. Telephone bandwidth

For evaluation of speech recognition using telephone bandwidths, the content of the database was filtered to the frequency range 300 Hz - 3400 Hz. The speech was also downsampled from 16 kHz to 8 kHz. In order to investigate the effect of the frequency range reduction, speaker-dependent recognition tests were performed using filtered data. All other settings were the same as in the test with full bandwidth. The average phoneme error was now 10.6 % for male speakers and 11.0 % for female speakers. The amount of correct phonemes was 92.0 % for both male and female speakers. All the following results in this article have been made by using this telephone bandwidth speech.

2.3. Speaker independent recognition

Speaker-independent speech recognition was experimented by training one recognizer with the speech of 9 male speakers. The training set consisted of the first speech sets. For the testing set consisting of the fourth speech set of the 9 male speakers, the average phoneme recognition error was 23.4 %. Due to the different speaking rates, this number was computed by changing all double phonemes to single phonemes. Similarly computed recognition error was 26.3 % for three male speakers who were not involved in the training. The amount of correct phonemes were 82.5 % and 80.9 %.

2.4. Speaker clustering

In order to improve the recognition rate, the speaker-space can be partitioned into overlapping cells and one speech recognizer can then be trained for each cell. But before one can find the speaker clusters, an appropriate distance measure between speakers must be defined.

The Self-Organizing Map was used as a codebook and pdf approximation for each phoneme. Let A_c and B_c be the codebooks for the phoneme c of two speakers. The following expression measures the amount of closeness of the pdfs of two speakers:

$$\sum_c P_c\left(\frac{\sum_i \min \mathrm{d}(A_c(i), B_c)}{\mathrm{size}(A_c)} + \frac{\sum_j \min \mathrm{d}(B_c(j), A_c)}{\mathrm{size}(B_c)}\right), \quad (1)$$

where index c goes through all phoneme classes, P_c is the corresponding a priori probability of the class and $\min \mathrm{d}(A_c(i), B_c)$ is the smallest distance from codevector i of codebook A_c to any codevector in codebook B_c by using the Euclidean distance. This measure may be denoted as a codebook distance. Probability functions may be compared by using distance measures such as Kullback-Leibler, Bhattacharyya or Matusita [2]. The

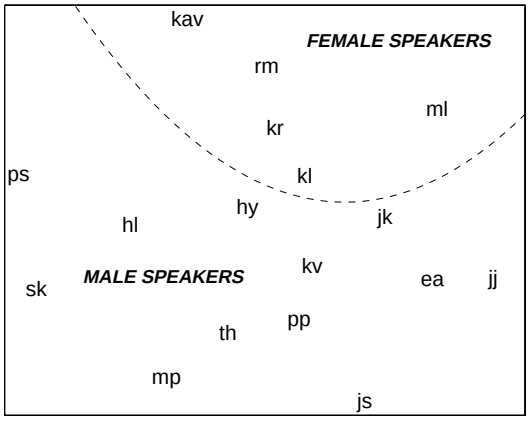

Figure 1: *Sammon mapping of speakers using codebook distances. The manually drawn dashed line shows that male speakers and female speakers are separated.*

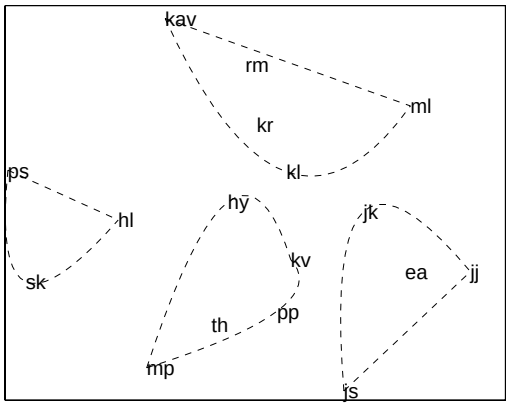

Figure 2: *One result of K-means clustering of the Sammon mapping of the speakers. Female speakers formed one cluster, which is the topmost cluster in the figure.*

distance measure used here resembles the Kolmogorov distance. Measure (1) is in fact a weighted sum of quantization errors of codebook pairs of all classes. The quantization error is computed in both directions (from codebook A_c to B_c and vice versa) in order to make the measure symmetric. Since the Self-Organizing Map is used as a codebook, the codevectors are located more densely in those areas of the feature space where the amount of the input data is high so that as a result each codebook approximates the pdf of its class. Although the measure (1) does not necessarily satisfy the triangle inequality needed for a definition of a proper metric, this measure was found useful in discriminating different speakers. When the speech of one speaker was recognized by using a speaker-dependent recognizer of other speakers, the recognition rate declined as the codebook distance increased.

One could argue that this speaker clustering is not appropriate because it is based only on static pdfs of the phonemes and does not take the transition probabilities of HMM states into account. However, the feature vector used here was a context vector which bears also information about the dynamic properties of speech.

In order to visualize the speaker clusters, the Sammon mapping [7] was used to represent the locations of the speakers in a two-dimensional plane. Sammon mapping is a method to project objects from a high-dimensional space into a lower-dimensional space while at the same time trying to preserve the relative locations of the objects. Here the Sammon mapping was computed according to the codebook distance (1).

Phoneme error% / correct%		
K	speaker pp	ea
1	18.6 / 89.6	22.4 / 86.6
3	20.8 / 92.0	22.6 / 88.7
5	21.9 / 91.6	24.7 / 87.9
7	21.2 / 91.1	18.6 / 90.7

Table 1: *Recognition rates by using several speech recognizers in parallel. K nearest speakers of the test speaker were selected according to the codebook distance and the corresponding speaker-dependently trained recognizers were used. The recognizer of the test speaker was not used.*

Speakers were then divided into four clusters using the K-means algorithm [5], and one speech recognizer was then trained for each cluster. When the recognizer was tested with those speakers involved in the training but using a different speech set, the average phoneme error rate was 19.5 %. For the speakers not involved in the training, the recognition rate was always the best with the recognizer belonging to the nearest cluster of a test speaker. But for the three speakers whose recognition rate was 26.3 % using the recognizer trained by 9 speakers, the recognition rate using cluster recognizers was 24.3 % and the phoneme errors didn't drop significantly. It can be seen that the number of speakers available in the speech corpus was too small in order to cover the speaker-space densely enough. The speakers were too disjoint in order to train a model which could perform well with all speakers not involved in the training.

In order to cover the speaker space, the speaker clusters could be made to overlap and thus compensate for the missing data of some speakers. Several speech recognizers can also be used in parallel and then form the final phoneme sequence by majority voting of the results of individual recognizers frame by frame. Table 1 shows an example of this.

3. EXPERIMENTS AT NRC

Two sets of experiments were carried out in the Speech and Audio Systems Laboratory, Nokia Research Center in Tampere, Finland. A subset of the database of the previous experiments, 14 speakers, two of them female, were used and models were trained with the standard HTK - Hidden Markov Model Toolkit [10]. First, 14 speaker-dependent systems were trained and evaluated, then a speaker-independent system was studied. The main goal of the development was to achieve competence in phoneme-based continuous speech recognition for Finnish applying hidden Markov modeling technique and to provide a solid basis for further research in continuous speech recognition over the telephone.

The most widely applied mel-scaled cepstral representation was used. Feature vectors were formed by appending the 0th cepstral coefficient values to the original 12 mel-cepstral coefficients along with their delta and delta-delta values. Only the 300-3400 Hz band was utilized for feature extraction with 25 ms Hamming-windows in 10 ms steps. Prior to the windowing pre-emphasis was applied. Some initial experiments indicated that neither mean subtraction, nor RASTA filtering yielded enhanced performance, therefore the above described simple front-end was used for the whole set of experiments.

The Finnish phoneme set to be modeled consists of 8 vowels ($/a/$, $/e/$, $/i/$, $/o/$, $/oe/$, $/u/$, $/y/$ and $/ae/$) and 12 consonants ($/d/$, $/h/$, $/j/$, $/k/$, $/l/$, $/m/$, $/n/$, $/p/$, $/r/$, $/s/$, $/t/$ and $/v/$). Additionally, a silence model was applied to represent short pauses between words. Models were chosen to be three-state left-to-right models, except for the silence model which was a one-state model. No skipping was allowed. The recognizer did not utilize any language model. Thus a simple phone-loop was applied.

3.1. Experiments and results

Two sets of experiments were carried out. First, speaker-dependent systems were trained and after evaluation the results were analyzed. The development of these 14 systems was carried out in successive steps applying multi-mixture Gaussians, introducing "double-phonemes" (see later) and experimenting with context-dependent models. The second set of the experiment concentrated on the development of a speaker-independent system. Results are presented in terms of phoneme correctness and error rates. The confidence interval at a 10 % error rate for the database in question is ca. ± 1 %.

3.1.1. Speaker-dependent system

The development of the 14 systems started with one-mixture monophone models. For each speaker a set of models was trained on the corresponding training sets. It was observed that one-mixture monophone models gave very poor performance. The average phoneme error rate achieved was higher than 40 %. As a second step, the number of mixtures in each state of each model of the 14 speakers was increased successively from 1 to 16. At the same time, the window sizes for the computation of the dynamic parameters were tuned individually for each speaker. The average window sizes for the delta and delta-delta parameters were 3.21 and 4.07, respectively. As a result, the performance of each system increased significantly. The mean error rate dropped to 11.9 %. The error rates varied between 9.0 % and 15.8 % for individuals.

3.1.2. Introducing "double phonemes"

It was detected that most of the errors occurred for long phonemes, that is when a phoneme occurred in the context of the same phoneme (e.g. in the word $saada$, $vaikka$). A closer look at the average duration of Finnish phonemes revealed that the Finnish long consonants and, especially, the vowels are extremely long in duration. For instance, while the phoneme $/a/$ lasts in average 146 ms, its long counterpart $/aa/$ is 319 ms on average. Therefore, it seemed to be a natural choice to introduce separate models for the long phonemes. The following "double phonemes" were added to the original set of phonemes: $/aa/$, $/ee/$, $/ii/$, $/oo/$, $/uu/$, $/yy/$, $/aeae/$, $/ll/$, $/mm/$, $/nn/$, $/pp/$, $/tt/$, $/kk/$). The number of emitting states varied between 6 and 10 for the "double phonemes" (vs. only 3 for the single phonemes). The total number of models thus increased to 34. This ad hoc duration modeling yielded a significant improvement for each system. The average phoneme correctness was 93.9 %, while the mean error rate decreased to 7.2 %. The individual error rates varied between 3.5 % and 11.9 %. On the word level, 67.6 % of the words were recognized correctly. Comparing the results to that of the systems with only single phonemes, the average error rate decreased by 40 %. This improvement is due to the more precise detection of long vowels, nasals, plosives and long $/l/$.

3.1.3. Context-dependent modeling

Up to now only monophone models were utilized in the 14 systems. In this experiment triphones were formed with the following scenario: after 5 embedded re-estimation cycles 1115 triphone models were created. 5 more re-estimation cycles were performed and the number of states of the 1115 models were reduced by state tying [11]. With this procedure the number

speaker	%SI_Corr	%SI_Error	%SD_Error
HL	94.9	14.8	4.5
SK	87.9	15.7	7.1
JK	90.2	14.7	4.8
TH	93.2	11.0	6.2
JJ	91.6	10.0	7.2
JS	91.4	13.1	8.2
PS	85.5	22.9	7.9
KV	79.0	23.9	9.6
MP	82.0	21.0	11.9
average	88.4	16.3	7.5

Table 2: *The performance of the speaker-independent (SI) system on a per speaker basis for the 9 training speakers.*

speaker	%SI_Corr	%SI_Error	%SD_Error
PP	91.1	13.70	3.5
EA	92.3	12.33	6.4
HY	77.1	26.55	11.2
KK	84.8	23.57	6.7
JH	81.5	22.48	12.0
average	85.3	19.7	8.0

Table 3: *The performance of the speaker-independent (SI) system on a per speaker basis for the 5 outliner speakers.*

of active states, and eventually the number of models can be decreased since similar items (in the statistical sense) are tied together. After state-tying, 18 additional re-estimation cycles were performed with the number of mixtures in each state of each model increasing from 1 to 8. The results indicate that no significant improvement was achieved. The mean correctness increased to 95.4 % and the average error rate decreased to 6.1 %. The error rates varied between 3.2 % and 9.5 %. It can be concluded that context-dependent modeling improved performance, however, only to a small extent. The achieved result, that is no significant overall improvement was gained with context-dependent modeling, implies that from the coarticulation point of view the Finnish language is not "problematic", as was expected in the beginning of the study. It very well might be adequate to model separately only the diphthongs and some vowel combinations with the phoneme /i/. This assumption requires further studies. The overall gain (performance improvement and increased computational complexity, 1115 vs. 31 models) achieved with the triphone system is minor and is very expensive due to the extremely large number of models.

3.1.4. Speaker-independent system

Finally, a speaker-independent monophone system was trained. The training material consisted of the first recording sessions from 9 male speakers (altogether 9×350 words). Two test sets were formed: the first set contained the latest recording sessions of the 9 training speakers (9×350 words), while the second set consisted of the latest recording sessions of 5 speakers who did not take part in the training (5×350 words). The results for the 9 speakers are tabulated in Table 2 and for the 5 outliner speakers in Table 3. For comparison, the speaker-dependent error rates (%SD_Error) of the monophone systems are also tabulated in the tables.

As could be expected the speaker-independent system did not perform as accurately as the individually trained speaker-dependent systems. The average error rate was doubled (16.3 % vs. 7.5 %) and the error rates increased further for the five outliner speakers who did not take part in the training.

4. DISCUSSION AND CONCLUSIONS

The above results, supporting earlier ones, justify that the Finnish language is ideal for phoneme-based continuous speech recognition. Baseline experiments showed error rates of about 10 % in speaker-dependent tasks. In further experiments ad hoc duration modeling of long phonemes yielded significantly better results (avg. 7.2 % error rate). Speaker-independent recognition did not meet the requirements of below 10-20 % errors but speaker clustering showed some promise to improve results. Speaker adaptation is probably needed for satisfactory results.

One has to take into account that the above results were achieved with a rather small database. More precise modeling of phonemes (especially in the context-dependent case) could be obtained with a larger, more carefully designed database. Further work is needed on better modeling of long phonemes (adequate duration modeling), as well as better improved strategies for speaker independency.

Based on this feasibility study we believe that useful and practical communication aids for the hearing impaired, based on Finnish phonemic recognition, can be developed. As the next step we are implementing an experimental system where STT and TTS conversions are combined in the telephone network in order to provide communication service between a selected set of normal telephone users and a set of deaf subjects using textual display terminals.

5. REFERENCES

1. Alarotu, N. et al. (1997) Applications for the Hearing-Impaired: Comprehension of Finnish Text with Phoneme Errors. In this Proceedings.
2. Ben-Bassat, M. (1982) Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation. In Krishnaiah, P.R. and Kanal, L.N., editors, Handbook of Statistics, volume 2, pp. 773-791. North-Holland.
3. Kohonen, T. (1995) Self-Organizing Maps. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg.
4. Kurimo, M. (1993) Using LVQ to enhance semi-continuous hidden Markov models for phonemes. Proc. of 3rd European Conf. on Speech Comm. and Tech. Berlin, Germany. vol 3. pp. 1731-1734.
5. Linde, Y., Buzo, A., Gray, R. (1980) An Algorithm for Vector Quantizer Design. IEEE Transactions on Communication, COM-28, Jan, pp. 84-95.
6. Rabiner, L.R., Wilpon, J.G., Juang, B.H. (1986) A segmental K-means training procedure for connected word recognition. AT&T Technical Journal, vol. 64, pp. 21-40.
7. Sammon Jr., J.W. (1969) A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, vol. C-18, no. 5, May, pp. 401-409.
8. Somervuo, P. (1996) Context vectors and multiple feature streams in speech recognition. Master's Thesis. Helsinki University of Technology.
9. 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.
10. Young, Steve J. et al. (1995). The HTK Book. Entropic Cambridge Research Laboratory. Cambridge, UK.
11. Young, Steve J. & Woodland, Phil (1993). The Use of State Tying in Continuous Speech Recognition. Proc. of 3rd European Conf. on Speech Comm. and Tech. Berlin, Germany. vol. 3. pp. 2203-2206.