Phoneme recognition for the hearing impaired

Blomberg, M.

journal: Proceedings of Fonetik, TMH-QPSR
volume: 44
number: 1
year: 2002
pages: 109-112

http://www.speech.kth.se/qpsr
Phoneme recognition for the hearing impaired

Mathias Johansson1, Mats Blomberg2, Kjell Elenius2, Lars-Erik Hoffsten3 and Anders Torberger3
1 Voxi AB, 2 Centre for Speech Technology, TMH, KTH, 3 Polycom Technologies AB

Abstract
This paper describes an automatic speech recognition system designed to investigate the use of phoneme recognition as a hearing aid in telephone communication. The system was tested in two experiments. The first involved 19 normal hearing subjects with a simulated severe hearing impairment. The second involved 5 hearing impaired subjects. In both studies we used a procedure called Speech Tracking to measure the effective communication speed between two persons. A substantial improvement was found in both cases.

Introduction
Today there are no alternatives to text telephones or relay services for hearing impaired people that fully incorporate the flexibility of the telephone. Automatic phoneme recognition supported by the human speech processing capacity could offer new telephone solutions for the hearing impaired. This paper reports on experiments with displaying recognized phoneme strings as an additional information source to the residual hearing of a listener. Part of this work has previously been reported in Johansson, 2002.

Background
As Automatic Speech Recognition technology matures, the range of possible applications increases. However, a domain and speaker independent system able to correctly decode human conversations into strings of words is not realistic today. The system we have in mind needs to be general. This is the main reason for experimenting with phoneme recognition. In our system the only linguistic constraints are the phoneme bigram weights. Thus, it relies heavily on the human user's language capability.

Another motivation for phoneme recognition is speed. The more complex the system, the longer the decoding time.

Telephone communication aids
Today hearing impaired people often need to depend on text telephones or relay services. Text telephones work well with experienced users. An obvious drawback is that it is necessary for both parties to have access to one. In relay services, a third party is listening in on the conversation, translating what is being said into text. This enables telephone communication when the text telephone is not a practical alternative. Having a third party involved naturally restricts privacy, however careful the relay service operator is about confidentiality.

Previous work
Experiments investigating the possibility of using phoneme recognition as a hearing aid have been conducted in Finland (Alarotu et al., 1997 and Karjalainen et al., 1997). In the first offline trials were performed concerning the readability of phoneme strings with simulated recognition errors. For isolated words the error rate could be up to 11% causing practically no degradation in comprehension. For sentences, with their additional contextual information, the phoneme error rate could be as high as 25%, without substantially reducing comprehension.

Finnish spelling is highly phonematic and Finnish speakers have no trouble reading the phonematic equivalents of words. For Swedish the acceptable error rates would probably be smaller since the orthography is less phonematic.

In Karjalainen et al., 1997, phoneme recognition experiments are described. Speaker dependent recognition resulted in an error rate of around 10%, which was proven to be sufficient in an aid. Speaker independent recognition resulted in error rates around 20%. This was deemed insufficient for the task. It was concluded that speaker adaptation was probably necessary to get adequate performance.

The phoneme recognizer
The recognizer in our experiment is based on HMMs trained using the HTK Toolkit (Young et al, 1999) with speech from the Swedish tele-
phone speech database SpeechDat (Elenius, 2000). The recognizer engine used is our in-house StarLite (Ström, 1996). Monophones were selected as recognition units in advance of triphones due to speed considerations. Phoneme bigrams were trained on phonetic transcriptions of Swedish newspaper text.

**Speaker adaptation**

The performance of our speaker independent system was too low to be useful for our task. Therefore we adapted the phoneme models to the target speaker using 70 sentences and 311 words read by him. A 17.8% error rate was achieved. This is in line with the results for additional contextual information in (Alarotu, 1997).

**Experiments**

We wanted to test whether our subjects could interpret recognized phoneme strings together with remnants of the acoustic signal. We used a procedure called Speech Tracking for measuring communication performance (see below). We performed 10 five minute Speech Tracking sessions with each subject, five without the aid interleaved with five with the aid. A sequence of 2 sessions forms a session pair. The first experiment was performed with normal hearing subjects and the second with hearing-impaired ones.

**The Speech Tracking Procedure**

The Speech Tracking Procedure (De Filippo et al, 1978 and Spens, 1995) has mainly been used to measure the lip reading capability of profoundly hearing impaired people, often with the purpose of evaluating different technical communication aids. In Speech Tracking the experiment leader (sender) reads a text aloud and the subject (receiver) tries to repeat what was said. If something is not understood the sender repeats the phrase or word where the error occurred. If the receiver does not understand what was said after a predefined number of repetitions (blocking), the phrase or word will be given in text so that the "conversation" may proceed without losing any context. The tracking score for each session is the number of words transmitted (or in some schemes correctly identified) divided by the time elapsed for each session.

**The text used**

The text used in this experiment was *Bröderna Lejonhjärta* by Astrid Lindgren. This text has been used extensively in earlier Speech Tracking experiments at the department.

**The experimental setup**

The system was installed on three computers: one for phoneme recognition, one for the Speech Tracking program, and one for distorting the speech signal to the normal hearing listeners in order to simulate a rather severe hearing loss. The idea was to remove spectral cues, such as formants, but keep amplitude and F0-information. The Speech Tracking program automatically calculated all the tracking scores and measures used (Gnosspelius and Spens, 1992). After initial studies we decided to give a word in clear text as soon as the subjects could not repeat it. The simulated hearing loss was very severe and requiring more than one attempt per blocked word was considered to be too frustrating for the subjects. This repair strategy was used for all experiments. One male speaker was the experiment leader in all sessions.

**The subjects**

The first experiment was performed with 19 normal hearing male and female subjects aged between 20 and 70. In the second experiment, 5 subjects with a real hearing impairment, aged between 30 and 65, listened to the undistorted telephone signal.

**Results for simulated hearing impairment**

The speech tracking score (L) yields a measure showing the average number of words communicated per minute (w.p.m.). Figure 1 shows the average tracking score per session.

![Figure 2. Tracking score (L) for hearing impaired subjects: average per session; with and without aid.](image_url)

The average tracking score for the five sessions without aid is 19.7 w.p.m. and 23.8 for the sessions with the aid. The average improvement
with the phoneme transcription aid is 20.9% ((23.8-19.7)/19.7). We do not reach the level, \( L=40 \), at which communication is agreed to flow satisfactorily (Spens, 2001).

The speech tracking procedure adopted in this work gives another measure which may shed some light on the cause of the fairly low tracking scores. The ceiling rate (\( L_c \)) gives the number of words per minute for the correctly understood turns in a session without any repetitions. This means that the phrase must have been accurately repeated by the receiver without any previous sender repetitions of any part of the phrase. Since phrases successfully repeated after the first turn are the only ones that are taken into account when calculating \( L_c \) this is a measure of how fast communication flows when there are no errors. In the sessions without aid \( L_c \) is on average 53.6. In the sessions with the aid \( L_c \) is only 34.5. The reason for this lower \( L_c \) is that with the aid the subjects spend more time to decode each turn even if they (thought they) heard what was said. Without the aid they immediately responded without seeking confirmation in the phoneme string. Although the aid slows down error-free communication, \( L \) and \( L_c \) tend to converge much more in the aided sessions compared to the unaided sessions. This is what should be expected if the aid is really an aid (Spens, 1995).

**Word measures for simulated hearing impairment**

The second measure we have used is the relation between the number of blocked words (\( \text{BlockW} \)) in a session and the total number of words (\( \text{TotW} \)) transmitted in a session. The smaller the ratio \( \text{BlockW}/\text{TotW} \) (henceforth called \( \text{blocking-ratio} \)) the better the subjects have understood what has been said. Table 1 shows the number of blocked words and the total number of words for both the aided and the unaided sessions averaged for all subjects. It clearly shows that the total number of words is higher and the number of blocked words is lower in the aided sessions as compared to the unaided.

Although \( L_c \) is on average higher for the unaided sessions, the total number of words transmitted is on average 16.9% lower than in the aided sessions. This is explained by the fact that when subjects were stuck on a word or phrase in a turn it was much harder to resolve the entire phrase if they did not have the phoneme string to help them. Often several words per blocked phrase had to be given in clear text in the sessions without aid. This may also be seen in the blocking-ratio in Table 1. This was not the case in the sessions with the aid. Thus, although the ceiling rate is higher in the unaided sessions the total number of words is significantly lower than in the aided sessions. In the unaided sessions more than 28% of the words were blocked and had to be given in clear text. In the aided sessions the same number was only 6.8% - a 76% error reduction.

**Table 1. Total average \( \text{TotW} \), \( \text{BlockW} \) and blocking-ratio for all normal hearing subjects and all sessions.**

<table>
<thead>
<tr>
<th></th>
<th>Without aid</th>
<th>With aid</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{BlockW} )</td>
<td>26.7</td>
<td>7.68</td>
</tr>
<tr>
<td>( \text{TotW} )</td>
<td>93.9</td>
<td>113</td>
</tr>
<tr>
<td>blocking-ratio</td>
<td>28.4%</td>
<td>6.8%</td>
</tr>
</tbody>
</table>

**Results for hearing impaired**

The second study was conducted to investigate how the phoneme aid would work with people with real hearing impairments. The speech tracking scores are shown in Figure 2. Again we see that the subjects perform better with the aid. The averages for \( L \) with and without the aid are 27.6 and 21.9 respectively – an average improvement of 26%.

**Figure 2. Tracking score (L) for hearing impaired subjects: average per session; with and without aid.**

The average tracking scores are higher than in the first study. In this study we are closer to \( L=40 \) but maybe not close enough. This is because the hearing impaired subjects “had better hearing” than the subjects with the simulated hearing loss. This is also reflected in \( L_c \): 34.9 with the aid and 32.7 without. Note that the \( L_c \) value is quite similar for the aided and the unaided sessions. In the first experiment, \( L_c \) was much higher in the unaided case. This may
probably also be explained by the fact that the subjects “heard better” in this study. Since the subjects were more confident, they did not wait for confirmation in the phoneme strings in the aided sessions if they (thought they) heard what was said. This is probably the reason why we do not find the relatively low value of Lc we saw in the first study in the aided case.

**Word measures for the hearing impaired**

The average BlockW, TotW and blocking ratio measures are given in Table 2. The average blocking-ratio for the aided sessions is 3.6% and 8.7% for the unaided – a 59% error reduction.

We also see that more words were consumed both with and without the aid as compared to the simulated hearing loss study. This is probably caused by the, on average, less severe hearing loss among the hearing impaired subjects.

**Table 2. Total average TotW, BlockW and blocking-ratio for all hearing impaired subjects and all sessions.**

<table>
<thead>
<tr>
<th>BlockW</th>
<th>TotW</th>
<th>blocking-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without aid</td>
<td>8.97</td>
<td>103.23</td>
</tr>
<tr>
<td>With aid</td>
<td>4.83</td>
<td>134.33</td>
</tr>
</tbody>
</table>

**Discussion**

The results so far indicate an improvement in understanding both in terms of tracking score and the number of blocked words. We have reported a 21% increase in L for a simulated hearing loss and a 26% increase for the hearing impaired. Further we have shown a decrease in the blocking-ratio in the aided sessions as compared to the unaided with 76% and 59% respectively. This seems promising. However, we are quite far from L=40, the level at which communication can be said to flow satisfactorily.

The blocking-ratio measure probably shows a more correct view of the difference between aided and unaided sessions. Since we chose to give each blocked word in the text to the user at the first repetition, a word was only blocked once per phrase. Thus, even if a phrase is hard to decode, it is fairly easy to consume it rapidly by getting each word in the phrase in clear text at each repetition. This, of course, holds for the aided sessions as well. But in the aided sessions much fewer words were blocked per phrase than in the unaided sessions. Therefore the results may be somewhat biased in favor of the unaided sessions if we look at the w.p.m. measure, L.

The results indicate a learning potential. The duration of the experiments was too short, however, to assess what level of proficiency a potential user may reach. Also, there are primarily three different aspects influencing the learning curves in the present system: continuously increasing context, improved ability to decode the distorted speech (in the first study) and, finally, increased phoneme reading ability. Prolonged experiments are necessary to shed more light on these matters.

**Acknowledgements**

This work was carried out at the Centre for Speech Technology (CTT), supported by Vinnova, KTH and participating Swedish companies and organizations.

We would like to thank Karl-Erik Spens for valuable discussions and Martin Dahlquist for his help with programming the simulation of the hearing impairment.

**References**


