A PHONEME RECOGNIZER FOR THE HEARING IMPAIRED

Mathias Johansson¹, Mats Blomberg², Kjell Elenius², Lars-Erik Hoffsten³ and Anders Torberger³

1Voxi AB, 2 Centre for Speech Technology, TMH, KTH, 3 Polycom Technologies AB
Stockholm, Sweden

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, which measures the effectiveness of the communication system for two persons. A substantial improvement was achieved in both cases.

1. INTRODUCTION

1.1. Background

As ASR technology matures, the range of possible applications increases. However, a domain and speaker independent system able to correctly decode all speech found in communication between people into strings of words is not realistic with the current state of technology. The system we have in mind needs to be general, however, which is the main reason for experimenting with phoneme recognition. In our system, the only linguistic constraints imposed on the acoustic hypotheses are the phoneme sequence probabilities, the phoneme bigram weights. Apart from this, the system does not incorporate any linguistic knowledge, since it is only intended to output a phoneme string representation of the input speech. 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.

1.2. 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. This excludes text telephone users from contacting the majority of the hearing population. 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. Most people would probably not resort to this alternative if they had a choice.

1.3. Previous work

Experiments investigating the possibility of using phoneme recognition as a hearing aid have been conducted in Finland [2, 3]. In [2], off-line trials were performed concerning the readability of phoneme strings with simulated recognition errors in order to determine an upper limit of the phoneme error rate for acceptable comprehension. 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 18%, and for dialogues it could be as high as 25%, without substantially reducing comprehension.

Finnish spelling is highly phonematic and Finnish speakers have no trouble reading the phonemic equivalents of the orthographic words. For Swedish the acceptable error rates would probably have to be smaller since Swedish orthography is less phonematic. On the other hand one could argue that phonemes give direct clues to how words are pronounced, which should reduce the influence of whether the spelling is phonematic or not. The amount of phonematic similarity to spelling might only mean that users would need more or less time getting used to reading what is normally heard.

In [3], phoneme recognition experiments are described. Speaker independent recognition resulted in an error rate of around 10%, which was proven to be sufficient for use in an aid for the hearing impaired. 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.

2. THE PHONEME RECOGNIZER

The recognizer in our experiment is based on HMMs trained using the HTK Toolkit [4] with speech from the Swedish 5000 speaker telephone speech database SpeechDat [5]. The recognizer engine used is our in-house StarLite [6]. Monophones were selected as recognition units in advance of triphones due to speed considerations. Phoneme bigrams were trained on automatically generated phonetic transcriptions of Swedish newspaper text.
2.1. Improving performance by speaker adaptation

The performance of our speaker independent system was too low for our task. Therefore we adapted the phoneme models to the target speaker (the experiment leader) using 70 sentences and 311 words. We investigated the MLLR and MAP adaptation procedures. Several experiments were conducted: MLLR with different number of regression classes, MAP with different scaling factors and a combination of MLLR followed by MAP. The lowest error rates, 17.8% for words and 30.0% for sentences, were achieved with MAP and a zero scaling factor, which is similar to the second with hearing-impaired subjects as described below.

3. EXPERIMENTS

We wanted to test whether a human listener actually could use the recognized phoneme strings together with remnants of the acoustic signal when interpreting speech in a telephone communication setting. We used a procedure called Speech Tracking for measuring communication performance (see below). We performed 10 five minute Speech Tracking sessions with each of the subjects, five without the aid interleaved with five with. A sequence of 2 sessions forms a session pair. The first experiment was performed with normal hearing subjects and the second with hearing-impaired subjects as described below.

3.1. The Speech Tracking Procedure

The Speech Tracking Procedure [7] 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 aloud from a manuscript 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 method has several advantages. It gives an easily understood measure of the receiver’s ability to comprehend what is being said (i.e. to track speech), and mimics, at least in part, a natural communication situation. It provides a convenient framework for a more or less controllable experiment on a very difficult task. It is also suitable in that no special equipment is needed and in that no complicated procedure has to be learnt or observed by the subjects.

The method also has several shortcomings. There are a number of parameters that may cause variability within a test set and, perhaps more importantly, when comparing different test sets. In the following we describe four parameters that may affect Speech Tracking results [8].

3.1.1. Text difficulty

This is quite obvious. The number of difficult or unusual words, the complexity of the sentence structure and the sentence length will all affect the receiver’s chance of successfully decoding what is said.

3.1.2. Speaking rate (sender and receiver)

In daily conversation the speaking rate is adjusted to a point where communication flows optimally. This also happens in the speech tracking sessions. Adjustment occurs more or less unconsciously and will naturally vary between speaker/listener pairs and therefore it will bias the tracking score.

3.1.3. Repair strategy

The repair strategy, i.e. the number of repetitions, the size of repeated units (word-, phrase- or sentence level) and the way the subjects are presented with a blocked word will also affect the communication speed and thus the tracking score. Therefore it should be well defined and used consistently throughout the entire experiment.

3.1.4. Certain language characteristics

Different languages differ in average word length. In English the average length of a word is 4.96, for Swedish 5.94 and for German 6.78 phonemes/word. This also influences the tracking score. A language with a higher average word length can be expected to give lower tracking scores.

3.2. 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. Although it is a book aimed at young readers, it is fairly difficult in that it contains a lot of names and in that the mode of expression is in some parts quite archaic.

3.3. 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 distortion was performed by an 8th order 250 Hz low pass filter and an amplitude dependent speech-weighted masking noise, which was added to the signal to keep the SNR constant. The idea was to remove spectral cues, such as formants, but keep amplitude and F0-information. This signal was picked up by a separate microphone for practical reasons. The signal to the recognizer came from a telephone over an ISDN connection.

The Speech Tracking program automatically calculated all the tracking scores and measures used [9].

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.

3.4. The subjects

The first experiment was performed with 19 normal hearing male and female subjects aged between 20 and 70. Nine had previous experience from reading phonetic transcriptions (trained subjects). The remaining 10 subjects had no such experience (untrained subjects). Note that by “trained” we do not mean trained in reading the particular phoneme strings output.
by a recognizer. Rather it means that the trained subjects were familiar with reading phonetic representations in general and that they had general knowledge of phonetics. It is uncertain to what extent this knowledge was an advantage in the task of fluently reading long phoneme strings and simultaneously adjusting for recognition errors.

In the second experiment, 5 subjects with a real hearing impairment, aged between 30 and 65, listened to the undistorted telephone signal. They were all “untrained”.

4. RESULTS FOR SIMULATED HEARING IMPAIRMENT

The conventional speech tracking score \( L \) is a simple metric computed as the total number of words transmitted per session divided by the total time of the session. This yields a measure showing the average number of words communicated per minute (w.p.m.). Figure 1 shows the average tracking score per session for all 19 subjects.

![Tracking score (L): average for all normal hearing subjects per session; with and without aid.](image)

The trained subjects performed slightly better than the untrained, but we would need more subjects in each group and more sessions for the results to be deemed significant.

4.1. 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 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 \( L \) 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>
<thead>
<tr>
<th>Without aid</th>
<th>With aid</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlockW</td>
<td>26.7</td>
</tr>
<tr>
<td>TotW</td>
<td>93.9</td>
</tr>
<tr>
<td>blocking-ratio</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

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

5. 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%.

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 \( L_c \) we saw in the first study in the aided case.
5.1. Word measures for the hearing impaired

The average BlockW, TotW and blocking ratio measures are given in Table 2. The relationship reported for the experiment with a simulated hearing loss holds here as well. The difference in blocking-ratio between the aided and the unaided sessions is not as dramatic however. The average blocking-ratio for the aided sessions is 3.6% and 8.7% for the unaided sessions – a 59% error reduction with the aid.

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

<table>
<thead>
<tr>
<th></th>
<th>Without aid</th>
<th>With aid</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlockW</td>
<td>8.97</td>
<td>4.83</td>
</tr>
<tr>
<td>TotW</td>
<td>103.23</td>
<td>134.33</td>
</tr>
<tr>
<td>blocking-ratio</td>
<td>8.7%</td>
<td>3.6%</td>
</tr>
</tbody>
</table>

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

6. 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 normal hearing subjects with a simulated hearing loss and a 26% increase in L for the hearing impaired subjects. 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 which would be the level at which communication can be said to flow satisfactorily. Here, one should keep in mind that the task was quite difficult. With the severe kind of hearing loss “simulated” for the normal hearing subjects, one would probably not use a standard telephone anyhow. Further, the text was quite difficult and quite different from how we speak.

The blocking-ratio measure probably shows a more correct view of the difference between the two groups of 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. This means that even if a phrase is very hard to decode, it is fairly easy to consume it rapidly by just giving 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. It is therefore difficult to estimate which of them has what impact on the results reported with the limited amount of sessions per subject at hand. Prolonged experiments are necessary to shed more light on these matters.

7. 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 for the normal hearing subjects.

8. REFERENCES