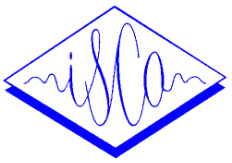

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

LEARNING AND TRANSFER OF LEARNING FOR SYNTHETIC SPEECH

Martine van Zundert, Jacques Terken

IPO, Center for User-System Interaction, Eindhoven, The Netherlands
{m.v.zundert, j.m.b.terken}@tue.nl

ABSTRACT

Understanding synthetic speech involves a learning process. We address the question whether transfer of learning takes place from one kind of synthetic speech to another. An experiment is presented in which learning curves for intelligibility were determined for two diphone-based synthesis systems for Dutch, A and B, with different diphone databases. Twenty-four subjects heard eight blocks of 50 Semantically Unpredictable Sentences (SUS). Four different experimental conditions were constructed: In conditions AA and BB, the same synthesis system (A and B, respectively) was presented in blocks 1 to 4 and 5 to 8. In conditions AB and BA one system was presented in blocks 1 to 4 and the other system in blocks 5 to 8. Results show that learning effects are observed within systems (conditions AA and BB). However, we find no evidence of transfer of learning between systems (conditions AB and BA).

1. INTRODUCTION

When we are confronted with speech of someone we don't know, we usually adapt quite quickly to the peculiarities of that speaker. The more the pronunciation of the speaker resembles what we are used to, the faster we adapt. For speech produced by non-native speakers or by speakers of a dialect we don't know, it may take quite a while to get used to and learn to understand it.

In this respect one might consider synthetic speech to constitute a particular dialect or non-native pronunciation. Indeed, it has been shown that understanding synthetic speech involves a learning process, and that the intelligibility of synthetic speech improves as a function of training/exposure (Schwab, Nusbaum & Pisoni, 1985; Francis & Nusbaum, 1998).

In this paper we focus on the nature of the knowledge that is learned: is the knowledge bound to the particular system by which it was obtained or is it sufficiently general so that it can be transferred to other systems as well. In order to investigate this issue, we ask whether training with one type of synthetic speech leads to an improvement in intelligibility for another type of synthetic speech.

The rationale of our approach is as follows. We determine improvements in intelligibility for two diphone-based synthesis systems for Dutch, A and B, with different diphone databases. Intelligibility is measured by means of presenting Semantically Unpredictable Sentences (SUS) in consecutive blocks of fifty utterances. Listeners receive feedback in order to facilitate training. Subjects are assigned to four different experimental conditions. In conditions AA and BB (the homogeneous conditions), the same synthesis system (A and B, respectively) is presented in blocks 1 to 4 and 5 to 8. In conditions AB and BA (the heterogeneous conditions), one system is presented in blocks 1 to 4 and the other system in blocks 5 to 8. To verify whether transfer occurs, we compare the intelligibility scores in blocks 5 of the heterogeneous conditions with those of blocks 1 in the homogeneous conditions. Also, we may compare the intelligibility scores in blocks 5 of the homogeneous and heterogeneous conditions to determine the amount of transfer.

2. METHOD

2.1. Materials

Experimental stimuli were produced by means of two diphone based text-to-speech synthesis systems for Dutch, which employ different diphone databases and different synthesis techniques. One system (further identified as system A) employs PSOLA-based synthesis, the other system (system B) employs MBROLA. Also, System A (PSOLA) has a female voice and system B (MBROLA) a male voice. The systems were shown to be approximately equal with respect to intelligibility in previous studies (Sluijter et al., 1998).

The SUS test was applied to measure the intelligibility of both synthesizers (Benoît, Grice & Hazan, 1996). SUS stands for Semantically Unpredictable Sentences. The SUS test is a standardised method for the assessment of intelligibility of synthetic speech at sentence level. The words that constitute a sentence are selected at random from a corpus on the basis of part-of-speech information only, so that the sentences are semantically anomalous. Thus, the listeners cannot compensate for poor quality of individual words by means of contextual knowledge. In addition, the words are rather short in order to avoid contributions from lexical knowledge. Finally, the sentences contain eight words or less, in order to avoid saturation of listeners' short-term memory. The syntactic structure of the sentences is grammatical and five simple different syntactic structures are used. Below, the five structures are shown with an example in Dutch and the literal English translation. The words printed in italics were given sentence accent in this experiment.

1. **intransitive structure**:
   Het *kind* rijst in het blonde *stuk*.
   (The child rises in the blond piece.)

2. **transitive structure**:
   Een Duits *bed* leidt het *plan*.
   (A German bed leads the plan.)

3. **imperative structure**:
   Neem nu de *wijn* en het *woord*.
   (Take now the wine and the word.)

4. **interrogative structure**:
   Hoe kreeg de *weg* de grijze *vriend*?
   (How did the road get the grey friend?)

5. **relative structure**:
   De *kant* heft de *staat* die *spijt*.
   (The side raises the state which regrets)

The text versions of the generated sentences were entered into the text-to-speech synthesis systems and converted into spoken text. Grapheme-to-phoneme conversion errors were corrected and sentence accents were assigned manually by editing the text.

2.1. Subjects

Twenty-four subjects (both male and female) participated in the experiment. All were native speakers of Dutch. They had no experience in listening to synthetic speech, and did not report hearing loss. The age of the listeners ranged from 18 to 24 years. All subjects were paid for participation. In addition, a bonus was given depending on their performance in order to encourage them to perform well.

2.3. Procedure

Subjects were seated in a sound-treated room in front of a personal computer. The procedure of the experiment was explained to the subjects in a written instruction. The actual experiment was preceded by a training session aiming to familiarize the subjects with the experimental task and with the type of sentences used in the experiment. In the training session subjects transliterated ten semantically unpredictable utterances (two instances of each of the five syntactic structures). In order to avoid training on synthetic speech the 10 semantically unpredictable sentences were recorded from a human speaker, a trained female speaker. Questions remaining after the training session were answered and then the experimental trials started.

At each trial, subjects were presented with an utterance and then entered what they thought the utterance had been into the computer. Subsequently, simultaneous visual and auditory feedback was given by displaying the correct transcription on the computer screen and playing the utterance. That is, the subject was not explicitly informed about whether his/her transcription had been right or wrong. Simultaneous visual and auditory presentation was supposed to enable the subject to learn the mapping between acoustic properties and phoneme labelling for the particular system. In all, subjects worked through eight blocks of 50 semantically unpredictable sentences. Presentation and pacing of the stimuli and data collection were under control of the computer. The stimuli were presented to the listeners by a closed headphone at a comfortable loudness level.

Four groups of 6 subjects were assigned to four different conditions. In conditions AA and BB, the same synthesis system (A and B, respectively) was presented in blocks 1 to 4 and 5 to 8. In conditions AB and BA one system was presented in blocks 1 to 4 and the other system in blocks 5 to 8. Utterances within each block were randomised for each separate subject. Likewise, the order of blocks was randomised in such a way that the same group of sentences would occur in different serial positions for different subjects. One block took about 15 minutes to finish and there was a coffee break of 15 minutes between groups of two blocks. The entire experiment took 4 hours.

2.4. Analysis

The standard procedure to compute an Intelligibility Index for Semantically Unpredictable Sentences, as described by Benoî t et al. (1996) is to calculate the percentage of utterances transcribed correctly. For an utterance to be counted as correct, all words in that utterance must have been transcribed correctly. However, this method only gives a rough estimate of the performance of the listener, since one incorrectly transcribed sound gives the same score for a particular utterance as three words transcribed fully incorrectly. In this experiment, a more precise way of scoring was considered desirable and for that reason we determined the score by the number of content words transcribed correctly in a block of 50 sentences. Function words such as determiners and prepositions were left out of consideration, since they are limited in number. Including these words in the score would inflate the performance, as subjects might easily guess them after a while. Transcription errors involving assimilation (*dampt* versus *damt*) were scored as correct, as well as homophones (*leidt* instead of *lijdt*).

In presenting and discussing the results, we denote each block in each condition by a code consisting of the condition (AA, BB, AB, BA) followed by the block number (1 to 8). So, the fifth block in condition AA is AA5, the first block in condition AB is AB1 etc.

3. RESULTS AND DISCUSSION

A score was computed for each block of 50 sentences and plotted in a curve (Figure 1 below). As can be seen, in all conditions blocks show an increase in percentage correct with an increase in serial position. To determine whether this progression is significant, analysis of variance (repeated measures on blocks, 'system' as between-subjects factor) was carried out on the data of conditions AA and BB. The effect of serial block position was significant ($F_{7,22} = 10.44$, $p \leq 0.005$) There was no difference in performance between the two systems used ($F_{1,22} = 0.07$, $p = 0.80$). The significant effect of serial position on subjects' scores shows us that training with synthetic speech helps to improve listeners' performance.

To see whether training with one system has an effect on the performance of the other system, we compare performance for blocks where there has been no prior training (AA1 and BB1) with the performance for blocks where a different system has been presented in de preceding blocks (BA5 and AB5 respectively). In this way we compare performance for blocks where there has been no prior training with the same system. The mean score for AA1 is 73.8% , the mean for BA5 this mean score is 76.4, a non-significant improvement ($t_5 = -0.68$, $p = 0.526$). The score of AB5 also improves as compared to BB1: from 73.3% to 78.6**%** , but again, this difference is not significant ($t_5 = -1.53$, $p = 0.187$). Comparison of AA1 and BB1 with AA5 and BB5 respectively, however, shows a significant improvement for both the AA and the BB conditions ($t_5 = -3.56$, $p = 0.016$ and $t_5 = -3.80$, $p = 0.013$ respectively). This suggests that staying in the same conditions (AA and BB) leads to a

significant improvement for block 5 compared to block 1, but when the system is changed at the fifth block (in conditions AB and BA) performance at block 5 drops to the same level as the starting level of the corresponding system. This suggests that there is no transfer of learning.

In the previous comparisons we compared blocks where there was no prior training with the same system. Either there was no training at all (blocks AA1 and BB1), or subjects were trained with a different system (BA5 and AB5). Strictly speaking, however, these comparisons contain confounding factors. The performance at block 5 may reflect fatigue effects. After all, by that time subjects have worked their way through 200 utterances. Alternatively, the performance at block 5 in conditions AB5 and BA5 may reflect at least some learning effect, *viz*. subjects may have learned to deal with the peculiarities of the task. And of course, the performance at block 5 may reflect both effects at the same time. To control these two factors (fatigue and task habituation) we also compare performance at the fifth block in the heterogeneous conditions (AB5 and BA5) with the performance at the same block in the homogeneous conditions (BB5 and AA5 respectively), thus controlling for the effects of fatigue and habituation at the same level. When the performance is significantly lower in the heterogeneous conditions, then this drop in performance must be due to the need to learn the properties of another system and not to any of the confounding factors.

For the comparison of AA5 with BA5, we find a significantly lower score for the BA condition ($t_5 = 3.09$, $p = 0.027$). This is in accordance with our conclusion above that there is no transfer of learning, since we find that what is learned for system B in condition BA does not generalise to system A. However, for the other comparison (BB5 with AB5) the difference is not significant ($t_5 = 0.46$, $p = 0.662$), which would suggest that in this case there is transfer of learning. A closer look at the curve for condition BB shows a major trough in performance at the fifth block (BB5), for which we have no explanation and which we therefore treat as anomalous. For that reason, we consider the significant difference between AA5 and BA5 as more reliable, and this comparison supports our earlier interpretation that there is no transfer of learning between the two systems.

4. CONCLUSION

We have addressed the question whether training with one type of synthetic speech improves the performance with another type of synthetic speech. Semantically Unpredictable Sentences were used to measure intelligibility. Four different experimental conditions were constructed: two homogeneous conditions (AA and BB), where subjects listened to the same system during eight blocks of fifty sentences each, and two heterogeneous conditions (AB and BA) where the synthesis system changed at the fifth block of sentences.

The hypothesis was that when there is transfer of learning between two systems, listeners should do better at block 5 in conditions where there has been prior training with a different system but not with the same system (BA5 and AB5), than at the first block in conditions where there also has been no prior training with the same system (AA1 and BB1). Also, the difference between the performance in the fifth blocks in the homogeneous condition (AA5 and BB5) and the heterogeneous condition (BA5 and AB5) should not be significant in case of transfer of learning.

Results show that subjects' performance increased significantly throughout the experiment in both homogeneous conditions and that there was no difference between the two synthesis systems used. Results do not provide support for the presence of transfer between the two synthesis systems, however, as training with another system did not improve performance as compared to no prior training. Thus, we conclude that listeners learn the specific properties of a given speech synthesis system rather than some knowledge that can be generalised across different types of synthetic speech.

5. REFERENCES

1. Benoît, C., Grice, M. & Hazan, V. "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences". *Speech Communication, 18*: 381 – 392, 1996

2. Francis, A.L. & Nusbaum, H.C. "Perceptual learning of synthetic speech". *Paper presented at the 136th Meeting of the Acoustical Society of America, Norfolk, VA*, Oct. 12-16, 1998

3. Schwab, E.C., Nusbaum, H.C. & Pisoni, D.B. "Some effects of training on the perception of synthetic speech". *Human Factors*, 27(4), 395-408, 1985.

4. Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A.A., Swerts, M.G.J. and Terken, J.M.B., "Evaluation of speech synthesis systems for Dutch in telecommunication applications." ***In:*** *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Blue Mountains, Australia*, November 26-29, 213-218, 1998

Figure 1: Percentage of content words transcribed correctly for each block in each of the four experimental conditions. Each block comprises 50 sentences and each condition involves six subjects. For AB and BA the transition occurred at block 5.