LEARNING EFFECTS FOR PHONETIC PROPERTIES OF 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

We address the question of what is learned while listening to synthetic speech produced by means of diphone-based synthesis. In standard diphone-based speech synthesis, the diphone database contains a single token for each phoneme transition. Learning may occur at different levels: listeners may learn the mapping between acoustic properties of particular diphones and their phonemic labelling; or they may learn phoneme models; or they may learn realisations of phonological features. Predictions of the different hypotheses are tested in an experiment in which we determine the improvement in intelligibility as a result of training for a specially constructed set of stimuli. The results force us to reject the hypothesis that listeners learn realisations of phonological features. However, they do not exclude the possibility that subjects learn phoneme models, although the results supporting this hypothesis do not reach significance.

1. INTRODUCTION

Listeners who have no prior experience with synthetic speech usually find it difficult to understand what is being said, while experienced listeners such as those working in the field of speech synthesis usually have little trouble. This implies that exposure results in learning. Indeed, it has been shown that training helps to improve listeners’ performance on the perception of synthetic speech (Francis & Nusbaum, 1998; Schwab, Pisoni & Nusbaum, 1985; Van Zundert & Terken, 2000).

We may question what is the source of this improvement. Diphone-based speech synthesis, which is currently the dominant approach and which is the focus of the current study, involves the concatenation of phoneme transitions (diphones) that are excised from natural speech and stored in a database. Usually, the database contains a single token for each diphone. As a consequence, there is little variability in the realisation of phonemes as compared to natural speech. In such systems, several explanations can account for the learning effect:

- Listeners may learn the mapping between the acoustic properties of the specific diphones and the phoneme labelling. For instance, they may learn to identify the acoustic information contained in a particular diphone {be} as the phoneme sequence /be/ (we will use curly brackets { } to denote a set of acoustic properties and slashes // to denote phoneme labels or sequences of phoneme labels).

stimulus was played again and the correct transcription shown. This should enable the subjects to establish a mapping between

- Alternatively, they may construct phoneme models on the basis of the acoustic information contained in the diphones that constitute tokens for the individual phonemes. For instance, exposure to the diphone {be} may lead listeners to learn about properties of the phoneme /b/ in this particular type of synthetic speech.

- Finally, they may learn parameter settings for distinctive features in this specific brand of synthetic speech. For instance, exposure to the diphone {be} may lead listeners to learn about the realisation of the voicing distinction in plosives in this particular brand of synthetic speech.

Under the first hypothesis, training with {be} will strengthen the mapping between the acoustic information contained in the diphone {be} and the labelling /be/, but nothing more than that. Thus training with {be} will result in an improved intelligibility for the particular diphone {be} but not for the diphone {bo} or {ba}, because their acoustic properties differ from those of {be}.

Under the second hypothesis, training with {be} will provide listeners with information about the realisation of the phoneme /b/ in this particular brand of synthetic speech, and enable them to construct a phoneme model for /b/. As a result, training with {be} will lead to an improvement for {be}, but also for {bo} and {ba}. Under the third hypothesis, training with {be} will provide listeners with information about parameter settings for distinctive phonological features such as voicing in plosives. According to this hypothesis, training with {be} will result in improved intelligibility not only for {be} and {bo}, but also for the voiced plosives {de} and {ge}. Below, an experiment is described to test the predictions derived from the different hypotheses by means of explicit training.

2. METHOD

2.1. Design

The experiment consisted of a pre-test, a training session and a post-test. Eight consonant-vowel (CV) diphones were selected from a diphone database for the training session. These will be referred to as the training stimuli. They were selected on the basis of the results of a study in which the segmental intelligibility of diphones had been evaluated (Van Zundert & Terken, unpublished data). Diphones were chosen with relatively low intelligibility scores, in order to leave room for improvement during training. Training was provided by means of feedback: after subjects had heard a particular stimulus and entered the transcription, the auditory the particular auditory pattern and the phonemic labelling. The subjects received no explicit feedback on whether their
transcription was right or wrong. The training stimuli were also presented in the pre-test and post-test, in order to evaluate improvements in intelligibility as a result of training.

In addition, a set of 16 control stimuli was selected. These were presented during the pre-test and the post-test but not during the training session. Eight of these stimuli (the same phoneme stimuli) contained the same consonants as the training stimuli but a different vowel (and thus concerned a different diphone), the other eight same feature contained different consonants but were taken such that they shared distinctive features with the training stimuli. The predictions of the different hypotheses might then be evaluated by comparing the intelligibility in the pre-test and the post-test for the training stimuli, the same phoneme stimuli and the same feature stimuli.

2.2. Materials

Stimuli were generated by means of a computer programme that takes strings containing phonetic representations and control codes for prosody as input. The pitch contour and the temporal properties were generated by rule. Stimuli were synthesised with a pitch accent. The speaking rate was set such that the duration of the stimuli would be similar to that of monosyllabic words spoken in isolation by human speakers.

In the previous subsection we have acted as if the stimuli consisted of isolated diphones, in order to simplify the presentation. In fact, the synthesis system cannot generate isolated diphones as output: a CV stimulus is created by concatenating a SILENCE-C\_ transition, a C\_V, diphone and a V\_, something transition, where “something” may be a silence or another phoneme (composed of a V\_ transition and a C\_ – SILENCE transition), thus creating a CVC stimulus (identical indices denote identical phonemes). In order to avoid undue repetition of the same stimulus during the training session, the latter option was chosen; that is, the target CV diphones were embedded into C1VC2 sequences, with different C2 for each repetition of C1V. A final C was also added to the stimuli in the pre-test and the post-test, so that subjects were presented only with CVC sequences. For a survey of the stimuli see Table 1 below. Furthermore, as the set of training stimuli was rather small and subjects might easily notice the repeated occurrence of the same C1V sequences, a set of dummy stimuli was added to disguise the purpose of the experiment.

2.3. Subjects

Forty subjects, both male and female, participated in the experiment. All subjects were native speakers of Dutch. None of them reported hearing loss. Subjects had no experience in listening to synthetic speech. Their ages ranged from 18 to 40 years. Subjects were paid for participation.

2.4. Procedure

3. RESULTS AND DISCUSSION
of 7.8% can be seen (from 53.4% to 61.2%). This difference is, however, not significant ($t_1 = -1.867, p = 0.104$) and therefore the first hypothesis is not supported. The first hypothesis holds that training will lead listeners to establish a connection between the acoustic properties of the diphone and the phonemic labelling. Hypotheses 2 and 3 can only hold when hypothesis 1 is also true, in other words: an improvement for the stimuli under hypothesis 2 should also involve an improvement for the stimuli under hypothesis 1. Therefore, if hypothesis 1 is rejected, hypotheses 2 and 3 are also rejected by implication.

A closer look at the scores for the individual training stimuli shows that the (non-significant) overall difference between pre- and post-test does not result from approximately equal differences for all individual stimuli. By itself, it is remarkable that we did not succeed in teaching our subjects the mapping between a bundle of acoustic properties and the phonemic labelling for the eight training stimuli, even though the stimuli were presented five times and subjects received feedback about the labelling. However, this does not concern us here, since we are interested in the source of learning, not in whether learning occurs. Therefore, in the following we will focus on those stimuli for which we observed a learning effect at all, since only for these stimuli we can evaluate the three hypotheses presented in the Introduction. That is, we will inspect more closely the results for the stimuli “fuum”, “riel” and “zjeun” which improve substantially after training (cf. Table 1). A t-test shows that the difference between pre- and post-test performance for these three stimuli is significant ($t_2 = -9.827, p = 0.01$).

When we look at the scores for the corresponding stimuli under hypothesis 2 (the same phoneme stimuli “faag”, “ruun” and “zjoei”), the scores for the stimuli in the post-test are substantially higher than those in the pre-test. Again, this difference is significant; $t_3 = -4.911, p = 0.04$). So, when subjects are trained with the training stimuli, their performance improves both for the training stimuli and the same phoneme stimuli. For the corresponding same feature stimuli “vuug”, “wieg” and “sjueg”, however, there is no (significant) improvement in performance. In other words, provided that subjects learn the properties of a certain stimulus from the training stimuli set, their performance for the same phoneme in another diphone improves as well, but the performance for a diphone sharing the same feature stimulus does not. From this, we tentatively conclude that listeners, when confronted with particular diphones, extract some notion of how particular phonemes sound in this brand of synthetic speech.

4. CONCLUSION

We addressed the question of what listeners exactly learn while listening to synthetic speech. The focus of our investigation was diphone speech. This type of synthetic speech is special, because for each Phoneme1-Phoneme2 transition the database contains precisely one element, so that most of the variability inherent to human speech is lost. When looking at the initial consonant of a special training set of CVC stimuli, three things could happen. First, listeners may learn the mapping between acoustic properties of particular diphones and their phonemic labelling (hypothesis 1). In case of any performance on the initial consonant in that specific CV context should improve, but the performance on different stimuli realising the same phoneme should not. Second, when listeners learn phoneme models, the initial consonant in both the specific CV context and other CV contexts should improve. Finally, when they learn realisations of phonological features, also the performance of other initial consonants sharing phonological features should improve (hypothesis 3).

We found that training does not necessarily lead to an improvement in performance for the initial consonant of all stimuli in the training set. In fact, only the initial consonants of a few of the training stimuli used in this experiment improved with training. Our first hypothesis (listeners learn the acoustic properties between the specific diphones and the phoneme labelling) is not supported as such. However there is evidence that for some of the stimuli, there was an effect of training. Possibly, for the other stimuli more training is needed.

For those stimuli under this hypothesis for which we do find an improvement in performance, the corresponding same phoneme stimuli improve as well. For the corresponding same feature sounds, however, no improvement in performance can be observed, thus leaving no evidence for hypothesis 3. This pattern of results suggests that listeners infer some notion of how particular phonemes sound in this brand of synthetic speech.

5. REFERENCES


---

1 An exception may be made for the recently developed corpus-based concatenative approaches in which utterances are composed from fragments taken from a corpus of natural speech.
<table>
<thead>
<tr>
<th>training stimuli (hypothesis 1)</th>
<th>control stimuli same phoneme stimuli (hypothesis 2)</th>
<th>control stimuli same feature stimuli (hypothesis 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stimuli</td>
<td>Scores</td>
<td>Δ</td>
</tr>
<tr>
<td>fuum</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td>riel</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>zjeun</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>zoel</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>sek</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>hieg</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sjeeck</td>
<td>32</td>
<td>30</td>
</tr>
</tbody>
</table>

**Table 1:** Experimental stimuli under each hypothesis with their corresponding scores in the pre-test and the post-test. Scores represent the number of subjects out of 40 correctly transcribing the initial consonant. The column “Δ” contains the difference score for each stimulus between the pre-test and the post-test.