COMPREHENSION OF ULTRA-FAST SPEECH – BLIND VS. "NORMALLY HEARING" PERSONS

Anja Moos and Jürgen Trouvain

Institute of Phonetics, Saarland University, Germany
{anmo, trouvain}@coli.uni-saarland.de

ABSTRACT

This study explores how much speech can be temporally compressed and still understood by blind people who have daily practice with speech synthesis vs. sighted persons without such training. Three text modes were generated (formant synthesis, natural speech with and without pauses). These texts were presented to sighted listeners at rates between 9-14 s/s and to blind listeners between 17-22 s/s. The removal of pauses in compressed natural speech shows significant benefits at only few speaking rates. Results also show that synthesis is understood worst by sighted but best by blind listeners. The fact that some of the blind still understood speech at 22 s/s reveals the enormous flexibility of the brain in speech perception during the processing of ultra-fast speech.

Keywords: speech rate, perception, compressed speech, speech synthesis.

1. INTRODUCTION

This study investigates how much speech can be temporally compressed and still understood by a) blind people who have daily practice with speech synthesis and b) sighted persons without such training. In addition, we look whether there are differences in comprehension between fast formant synthesized speech and compressed natural speech. A third aspect investigated is whether the removal of silent pauses affects understanding.

In a study with synthetic speech [9], two blind students were able to understand synthetic speech at a speaking rate (tempo including pauses) of 17 syllables per second (henceforth s/s) whereas the comprehension of their non-blind peers falls off drastically at 9 s/s. That means that the non-blind were still able to follow the message at a tempo which corresponds to the most extreme rates of human speech production, whereas the blind subjects were able to go well beyond this point.

Interestingly, this result holds true for synthetic speech generated with a *formant* synthesizer. At normal rates (between 3 s/s and 5 s/s), diphone synthesis is generally considered more natural than formant synthesis and is therefore preferred. Fast speech at rates higher than 10 s/s produced with *diphone* synthesis was nearly as unintelligible for the blind as for the sighted persons. It is unclear how temporally compressed *natural* speech can be understood by the two groups.

Fast *human* speech features shorter segment durations, elisions, assimilations and reductions of segments, fewer pitch accents as well as fewer and shorter pauses. At all levels the changes occur *non-linearly* (cf. [8]). Janse [3] could show that speech at extreme rates (8.5 and 10.5 s/s) is *less* intelligible when it is modelled to natural fast hypo-speech rather than hyper-speech. She also found that "the only nonlinear aspect of natural fast speech that does improve intelligibility over strictly linear compression is pause removal. Note, however, that this only becomes advantageous when compression rates are relatively high" [3, p. 163].

The present study investigates the comprehensibility of various types of compressed speech (formant synthesis, natural speech with and without silences) for two groups of listeners (blind daily users of synthesis, sighted persons) at different speaking rates as illustrated in Figure 1.

Figure 1: Speaking rates at usual reading speed (left), ultra-fast for seeing persons (middle) and synthesis experienced blind persons (right).

```
||||||||           |||||||||||||||||||||||||||||             |||||||||||||||||||||||||||||||
====================================> s/s
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
```

2. METHOD

2.1. Subjects

21 blind persons participated in the study. Their age ranged from 15 to 37 (mean: 22 years); 13 of them had been blind from birth; 76% were males. All blind subjects had had daily experience with a formant synthesiser for more than two years (mean: 7 years).

For the sighted group, 21 persons with (on average) little experience with synthetic speech participated. Their age ranged from 21 to 34 (mean: 26 years); 52% were males.

All subjects speak German as their native tongue. Two blind and two seeing subjects were excluded because of slight hearing deficiencies. In total, the answers of 19 blind and 19 sighted persons were analysed.

2.2. Text material

To simulate a realistic screen-reader situation, authentic (German) texts rather than single sentences were used. They comprised 18 short texts (e-mails, informative texts and news) with a mean length of 102 words (standard dev.: 4 words) and 209 phonological syllables (sd: 32), respectively.

2.3. Stimuli

2.3.1. Baseline versions

Synthetic and natural speech was used. For both methods, first baseline versions at normal speech rates were generated or recorded, respectively. The baseline versions for the synthetic stimuli were generated with the formant synthesizer Eloquence [2] in the screen-reader software JAWS [4]. The baseline versions for natural speech were spoken by a professional speaker who was recorded in a sound-treated room reading the 18 texts at a self-selected speed. His speaking rates (including pauses) for the texts lie between 3.9 s/s and 4.5 s/s.

In the screen-reader software, the texts were generated with silent pauses at unacceptable locations. But, contrary to linguistic experience, there were no pauses at the end of sentences. We decided to cut out also the pauses at the unusual places. Consequently, the baseline versions for the synthetic stimuli (speech mode SYN) do not contain any pauses. A further feature that the blind often select is the mode "read with some punctuation marks pronounced", with the consequence that the synthetic baseline versions have more syllables than the natural ones.

For the natural speech we have two sets of baseline versions: the recorded one without any manipulations (INCL) and a second one with all silent and breath pauses carefully cut out (EXCL).

2.3.2. Compressed versions

The baseline versions were manipulated with the standard speech editor Praat [5] which makes use of PSOLA [1] as manipulation method: the fundamental frequency stays constant and the durations are changed linearly by averaging adjacent F0 periods which overlap in the time domain.

For each group of subjects, stimuli at six different speaking rates, 1 s/s apart, were produced. Sighted people were presented with stimuli at rates from 9 to 14 s/s; the stimuli for the blind ranged from 17 to 22 s/s. For each tempo, a subject heard a different text for each speech mode. Each subject judged 18 stimulus texts (3 modes x 6 tempo steps) in total. (Compare accompanying audio files with the same text in different modes at 14 s/s.) To minimize the effect that a given text influences the judgement of a given rate, the speaking rate of each text was rotated among subjects.

No matter whether there are silences or spoken punctuation marks, stimuli of the same tempo category have exactly the same speaking rate though not the same articulation rate. The example in Table 1 illustrates that stimulus duration can differ between SYN and natural-based versions, and that the articulation rate is higher in versions with silences (INCL) than without (EXCL).

Table 2: Example for one text at the same speaking rate in different modes: speaking rate (SR) including pauses, articulation rate (AR) excluding pauses, number of syllables and stimulus duration (in secs).

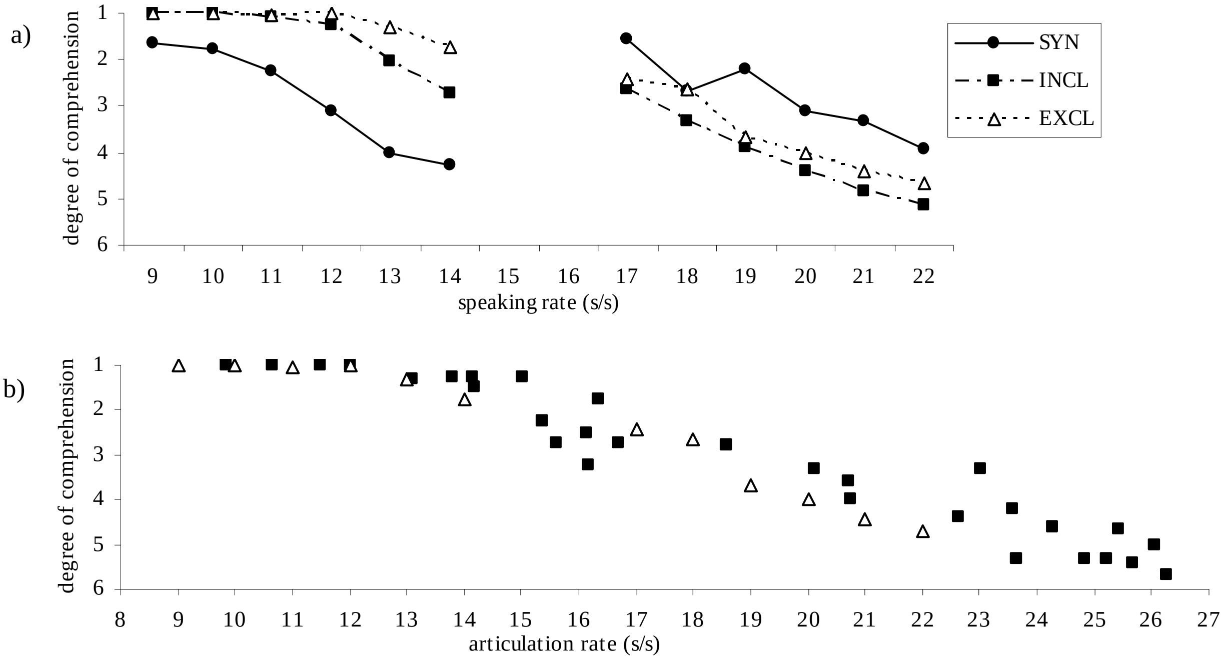
mode	SR (s/s)	AR (s/s)	# syll	stim dur
SYN	11.0	11.0	209	19.000
EXCL	11.0	11.0	196	17.818
INCL	11.0	13.2	196	17.818

2.4. Listening test

The test was performed in quiet rooms and started with questions as to age and educational background for both groups, duration of blindness and length of experience with formant synthesis for the blind, and the general experience with speech synthesis for the sighted. The texts were presented to the subjects via headphones connected to a laptop. Before listening, the subjects were instructed about the judgment procedure: after listening to a stimulus they had to give their subjective judgement on a six-step scale of how much they understood of the text. The six degrees were defined as: (1) "all", (2) "nearly all", (3) "more than half", (4) "less than half", (5) "nearly nothing", (6) "nothing".

Subjective comprehension rather than recall of content words was used because we were dealing

Figure 2: Degree of understanding as a function of speaking rate (top) and articulation rate (bottom) for compressed natural speech including silences (INCL), excluding silences (EXCL) and for formant synthesized speech (SYN). Sighted: 9-14 s/s, blind: 17-22 s/s.



with texts, not with sentences in a daily reading situation for users of screen-readers. However, it cannot be ruled out that some listeners scan rather than really comprehend the texts.

In [10] it was found that comprehension of compressed speech increased significantly over the first ten minutes with little increase after that. Thus, in the first ten minutes all subjects were "trained" on ultra-fast speech rather than just "warmed up" to the test condition. The tempo in these training files was comparable to the speaking rates of the stimuli presented in randomised order in the test phase.

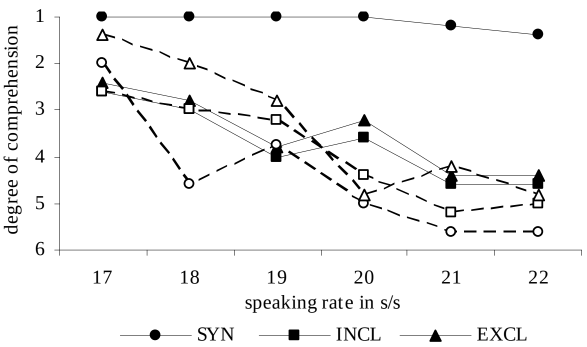
3. RESULTS

Figure 2 shows that the comprehension of the human based speech declines continuously with speech and articulation rate, irrespective of the group. For SYN, there is a separate decline for the sighted and for the blind since the blind are much better in understanding synthetic speech. For sighted persons, SYN is least understandable whereas for the blind it is understood best. Significant differences are at p<0.05 (Mann-Whitney-U-Test) between SYN and INCL/ EXCL at all speech rates for the seeing and 17 and 19 s/s for the blind as well as between SYN and INCL at 20 and 21 s/s.

There are significant differences between INCL and EXCL only at 13 and 14 s/s at p<0.05 (Mann-Whitney-U-Test) due to the deletion of silent and breath intervals. It can be seen in Figure 2b that the level of understanding is more strongly correlated with *articulation* rate (excluding pauses) rather than with *speaking* rate (including pauses). Differences in the comprehension within one tempo category between INCL und EXCL are therefore only due to articulation rate.

As the variance was larger among the blind in all modes, subgroups were created. Figure 3

Figure 3: The mean values for two sub-groups of the blind: the five best synthesis scorers (filled symbols/solid lines) and the five worst synthesis scorers (empty symbols/dashed lines).



compares the blind subjects with the five best versus the five worst results for SYN. The ability to comprehend ultra-fast synthetic speech is not transferred to the comprehension of ultra-fast compressed natural speech as five best scorers clearly show.

Interestingly, congenitally blind subjects had significantly poorer performance in the SYN condition than those who became blind after birth (Spearman Test: r=-0.527, p=0.02). Whereas subjects blind from birth had bad scores for SYN and medium ones for EXCL, subjects who lost sight in their childhood were best SYN scorers.

4. DISCUSSION

The degree of understanding of compressed natural speech decreases as its tempo increases for the ratings of sighted as well as for blind subjects. Although it can be expected that the blind would score slightly better than the sighted persons for the rates between 9 and 14 s/s, the extra-ordinary listening skills found for formant synthetic speech is not found for natural speech in very fast modes.

The deletion of silent and breath intervals does hardly affect comprehension. "Pauses" at these rates can still be perceived as prosodic breaks without having silent intervals because of final lengthening and falling intonation.

One explanation for the good results of formant synthesis for the blind is that all of our subjects listen to this sort of speech daily for a considerable period of their life. The results show how flexible the perception mechanism for speech can be after a long and intensive training. A longitudinal study of the training effect would tell us more about how we can learn to exploit this enormous flexibility. Training effect reaches a plateau after 10 minutes for compressed natural speech [10] or after 5 days for synthesised sentences [6]. The longitudinal effect of years of training to ultra-fast speech is largely unexplored.

An advantage for a better comprehension of formant synthesized speech is that it is very clear and sounds hyper-articulated (cf. [3]). In contrast to clear *artificial* speech, compressed clear *natural* speech – even if produced by a professional speaker – shows typical characteristics of hypo-articulation and a stronger coarticulation in unstressed syllables (i.e. in the vast majority of syllables in German). Obviously, redundant hyper-speech phenomena at normal rates are helpful for the understanding of highly compressed speech – but only for some of the trained listeners (Fig. 3).

In contrast to natural speech or synthesis with concatenated pre-recorded speech (e.g. diphone speech or unit-selection), formant synthesis shows no human voice quality. Usually, humans process all para-linguistic and extra-linguistic information (e.g. who is speaking) during listening. In formant synthesis this sort of processing may be switched off – if you are trained to do so.

The visual cortex plays a role for the congenitally blind for linguistic tasks such as complex syntax and word meanings [7]. The visual cortex might also be necessary for phonetic understanding (compare McGurk effect). It is, however, unclear if the difference found between congenitally and non-congenitally blind subjects is neurologically based. This hypothesis will be a topic for future neurophonetic studies. We hope that this study helps to explore and to explain the extra-ordinary perception skills during the processing of ultra-fast speech.

5. ACKNOWLEDGEMENTS

Many thanks to our reviewers and to our subjects, especially those from Carl-Strehl-Schule (Marburg).

6. REFERENCES

[1] Charpentier, F., Moulines, E. 1989. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Proc. *Eurospeech* (2), 13-19.

[2] Hertz, S. 1997. *The ETI-Eloquence Text-to-Speech System.* Eloquent Technology Inc, Ithaca.

[3] Janse, E. 2003. *Production and Perception of Fast Speech*. Lot, Utrecht.

[4] JAWS (Job Access With Speech) Screenreader software http://www.freedomsci.de visited 21-Jan-06.

[5] PRAAT version 4.5 http://www.fon.hum.uva.nl/praat/ visited 10-Jan-07.

[6] Reynolds, M.E., Isaacs-Duvall, C., Haddox, M.L. 2002. A comparison of learning curves in natural and synthesized speech comprehension. *Journal of Speech, Language and Hearing Research* 45, 802-810.

[7] Röder, B., Stock, O., Bien, S., Neville, H., Rösler, F. 2002. Speech processing activates visual cortex in congenitally blind humans. *European Journal of Neuroscience* 16, 930-936.

[8] Trouvain, J. 2004. *Tempo Variation in Speech Production. Implications for Speech Synthesis.* Phonus 8, Reports in Phonetics, Saarland University, Saarbrücken.

[9] Trouvain, J. 2007. On the comprehension of extremely fast synthetic speech. *Saarland Working Papers in Linguistics* (SWPL) 1.

[10] Voor, J.B., Miller, J.M. 1965. The effect of practice upon the comprehension of time-compressed speech. *Speech Monographs* 32, 454-456.