FAMILIAR TALKER ADVANTAGES IN FORMANT-BASED AND CONCATENATIVE SYNTHETIC SPEECH

Jacqueline Jones

Dept. of Linguistics, University of Calgary, 2500 University Drive NW, Alberta, Canada, T2N 1N4, jmjone@ucalgary.ca

1. INTRODUCTION

Access to synthetic speech technology has never been easier than it is today. Home computers come bundled with text-to-speech software, as do some eReaders and smart phones. The technology has come a long way since Stephen Hawking's recognizable DECTalk voice in the late 1980s. And yet despite synthetic speech's increased intelligibility, decreased cost, and lessened reliance on bulky equipment, there is still a recognizable peculiarity in synthetic speech. While extensive research has been done on both synthetic speech perception and familiarity effects, no study has yet examined whether synthetic voices are treated as individual voices, or whether they are perceived by the listener as a broad category of "synthetic speaker" that encompasses all voices produced in the same manner.

It has been established that listener perception of synthetic speech improves with training, but past research has also identified limits to this improvement (Nygaard et al, 1998; Greenspan et al, 1988). The perceptual benefits listeners gain from being familiar with a speaker in natural speech is known as the "Familiar Talker Advantage."

This research examines this familiar talker advantage and how it relates to synthetic speech by utilizing two kinds of synthetic speech. In this study, listeners were trained to identify words produced by four different synthetic voices, created using two different synthesizing processes.

2. METHOD

Participants were trained with a synthetic speaker produced by one of two types of synthesis (formant-based and concatenative) via a series of sentence transcription tasks. Participants were then tested on their ability to transcribe sentences produced by novel synthetic voices to determine if the training generalized across speakers and synthesis types. Finally, participants completed a post-test phase to examine the retention effects of any benefits from training.

2.1 Participants

24 young adult monolingual speakers of Canadian English participated in these experiments. Two participants were removed from the results for problematic answers or manipulation of testing equipment during testing. All participants were recruited from the University of Calgary's introduction to linguistics class in exchange for marks towards their research participation requirement for that class. Prior to starting the research, participants were asked to confirm that they were native English speakers and asked to provide a self-assessment of their familiarity with synthetic speech. None reported any speech or learning deficit, and all reported having normal hearing and minimal exposure to synthetic speech.

2.2 Stimuli

Four sets of stimuli were created for this research. The threshold stimuli consisted of 70 pre-recorded spondees – bisyllabic words with equal stress on both syllables – produced by a native speaker of American English. Training, Testing, and Post-Test stimuli consisted of pre-recorded sets of Harvard Sentences produced by synthetic speech. The formant voices were produced from presets in the eSpeak text-to-speech freeware program, version 1.45.05. The concatenative voices used were from the Ivona Voices commercial software suite. The names chosen from the Ivona Library were Joey and Eric.

2.3 Procedure

Participants were placed randomly into four groups.

During the threshold phase, all four groups listened to 70 pre-recorded spondees produced by a natural speaker of American English. The volume the tokens were played at was progressively reduced. The volume at which participants were achieving correct answers 50% of the time was used for all further phases of the experiment.

The training phase consisted of 60 sentences produced by a synthetic speaker. Groups 1 and 2 trained with the concatenative voice Eric, and groups 3 and 4 trained with formant voice Wheatley. The participants listened to the sentence a single time and were asked to transcribe what they heard. The sentence was then repeated, and feedback was given in the form of a transcription displayed on screen. The response and response time were recorded.

In the test phase, immediately after the training phase, participants were asked to transcribe 20 Harvard sentences. For two groups, 10 of the sentences were produced using the same synthetic speaker as presented in training, while 10 were produced by a voice which used a different type of synthesis. The other two groups were tested with 10 sentences presented by a voice that differed in synthesis type, and 10 sentences produced by a different synthetic speaker that shared a synthesis type with the voice used in training.

The post-test phase took place between three and five days after the initial training and testing. In the post-test phase, participants were presented with 20 novel Harvard sentences, of the same type and distribution as those used in the test phase.

3. RESULTS

Three ANOVAs were run to determine what effects trained voice, stimulus type, and test number had on the percentage of unique content and function words correctly identified, and response times. Post-hoc t-testing between individual pairs of variables was used to identify the directions of the significant interactions. Please contact the author for the complete results of the experiment.

Overall, participants identified individual words in testing correctly significantly ($p < 0.001$) more often when trained with the Formant-based synthesized voice Wheatley, regardless of stimulus type or the properties of the individual words (content or function).

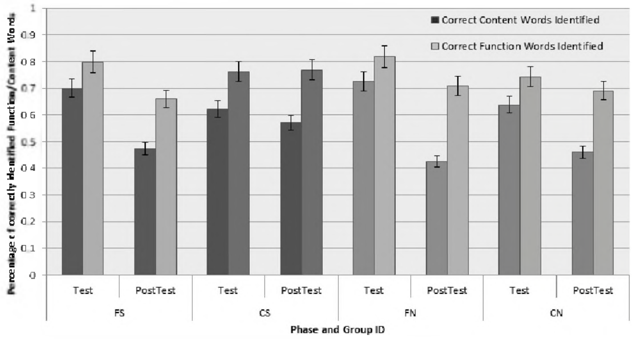

Figure 1. Word Identification scores across groups and phases.

Figure 1 shows word identification scores across all groups during the testing and post-testing phase. Darker colours indicate content word identification and lighter colours represent function word identification. Overall performance is worst during the post-test across all groups, and performance by those trained with formant-based voices is significantly better ($p < .001$). Those trained with formant voices do worse in the post-test than their counterparts trained with concatenative voices ($p < .0001$).

Function words were correctly identified more often regardless of any other factor. In examining individual results, when function words were incorrect, they were often transcribed as other function words – 'a' transcribed as 'the', for example.

Overwhelmingly, participants performed better in test one than in test two. Further refinement of the methodology is necessary to determine whether this drop in performance is a significant indication of failure to retain benefits gained in training over time.

4. DISCUSSION

The results of this research show support for the existence of a familiar talker advantage in synthetic speech. The results also show evidence of the existence of a familiar synthesis type advantage. The group trained with Eric and tested with Eric and Joey had better performance with the similar voice (Joey) than with the voice they were trained with (Eric). This could indicate a between-type perceptual benefit to concatenative synthesis. However, it is not possible to determine the extent to which this performance boost is due to a between-type perceptual benefit, a combination of Joey's novelty causing an increase in attention of the participant, or even whether Joey is perceived as a naturally more intelligible voice.

The results suggest that listeners gain a perceptual benefit from familiar synthetic talkers, regardless of synthesis type, in a similar way to the benefit gained from natural familiar talkers. Training with a concatenative synthetic speaker can potentially grant a "synthesis type familiarity" advantage, but the exact nature of this advantage cannot be determined from the current results, as there are too many variables between the acoustic properties of the two synthesis types to confidently determine which are causing the perceptual benefit.

Participants trained with the less intelligible formant-based synthesis voice had overall better performance with all synthesized voices presented to them. A perceptual benefit appears to emerge across synthesis types, but only in one direction and if the training is relatively intense. This may have an analogue in the research done on bilingual talkers in English and German, where listeners had more success with unfamiliar languages. (Winters et al, 2008). Effectively, this means that the more intelligible synthesized voices do not transfer any kind of familiar talker advantage to less intelligible voice, but that training with less intelligible voices may grant a perceptual benefit to synthetic speech in general.

REFERENCES

Greenspan, S. L. et al (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology*. 14, 421-433.

Nygaard, L. et al (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*. 60, 355-376.

Winters, S. J., et al (2008). Identification and Discrimination of Bilingual Talkers across Languages. *The Journal of the Acoustical Society of America*. 123, 4524.

AUTHOR NOTES

This research was conducted while the author was an undergraduate student at the University of Calgary. The author can be contacted at jmjone@ucalgary.ca.