

On the Intelligibility of Fast Synthesized Speech for Individuals with Early-Onset Blindness

ABSTRACT

People with visual disabilities increasingly use text-to-speech synthesis as a primary output modality for interaction with computers. Surprisingly, there have been no systematic comparisons of the performance of different text-to-speech systems for this user population. In this paper we report the results of a pilot experiment on the intelligibility of fast synthesized speech for individuals with early-onset blindness. Using an open-response recall task, we collected data on four synthesis systems representing two major approaches to text-to-speech synthesis: formant-based synthesis and concatenative unit selection synthesis. We found a significant effect of speaking rate on intelligibility of synthesized speech, and a trend towards significance for synthesizer type. In post-hoc analyses, we found that participant-related factors, including age and familiarity with a synthesizer and voice, also affect intelligibility of fast synthesized speech.

Categories and Subject Descriptors

K.4.2 [Social Issues]: Assistive technologies for persons with disabilities; I.2.7 [Natural Language Processing]: Speech recognition and synthesis

General Terms

text-to-speech synthesis, fast speech

1. INTRODUCTION

People with visual disabilities increasingly use text-to-speech synthesis rather than Braille as a primary output modality for interaction with computers. In fact, they probably represent the second largest user population (after developers of interactive voice-response systems) for text-to-speech synthesis. As they become expert users of synthesized speech, they may listen to the speech at speeds multiple times real time [2]. Consequently, they may have different performance metrics for text-to-speech systems than other user populations (e.g. they tend to prefer intelligibility over naturalness). They tend to have strong preferences for a particular

synthesizer and voice¹, based partly on familiarity but also perhaps on suitability of the engine or voice for their needs. Although some studies have analyzed the intelligibility of fast speech in general (e.g. [11, 12, 14, 15, 20]), there have been no comprehensive comparisons of the performance of text-to-speech systems for people with visual disabilities.

In this paper we report the results of a pilot experiment evaluating, for individuals with early-onset blindness, the intelligibility of synthesized speech across text-to-speech synthesis engines, voices and approaches. The goals of our research are to:

- Empirically identify the best text-to-speech engines for people with visual disabilities who listen to fast synthesized speech.
- Identify whether gender of the voice affects intelligibility of fast synthesized speech.
- Identify whether the method of synthesizing speech affects intelligibility of fast synthesized speech.
- Identify good metrics for evaluating the quality of text-to-speech synthesis for people with visual disabilities.

We hope that the answers to these questions will help drive research on text-to-speech synthesis for this user population.

The rest of this paper is structured as follows: In Section 2, we summarize the findings of previous work on the intelligibility of fast synthesized speech. In Section 3, we present the design of our pilot study, and in Section 4 we present the results. In Section 5 we conclude and summarize our current work and ideas for future research in this area.

2. RELATED WORK ON FAST SPEECH

Research on the intelligibility of fast speech in individuals with normal vision has shown that:

- Natural speech is more intelligible than synthesized speech [18].

¹See, for instance, the blog post by Sean Randall titled “Being eloquent about Eloquence: a formant toast”, published January 31, 2009 at <http://randylaptop.com/blog/being-eloquent-about-eloquence-a-formant-toast/>.

Personal profile

Fill in the form below and click on **OK**.

Your personal information will be treated as confidential.

In which age range are you?

Are you a native speaker of English?

If you are not a native speaker of English, what is your native language?

If you are not a native speaker of English, at what age (in years) did you learn English?

What do you listen to your computer on?

How often do you hear TTS speech?

Which TTS systems do you most often use?

Which kind of TTS voice do you prefer?

How do you change TTS output speed?

Figure 1: Demographic survey, including questions about participant age, English speaking ability, and experience with text-to-speech synthesis systems

- Linearly time compressed synthesized speech is more intelligible than both uncompressed fast natural speech and ‘naturally compressed’ synthesized speech [11] (cf. [15], who used subjective judgments rather than recall to evaluate intelligibility).
- Linearly time compressed natural speech is more intelligible than linearly time compressed synthesized speech [14].
- The intelligibility of fast speech can be affected by listener-related factors including age, hearing ability, language fluency, and familiarity with the synthesis engine and voice [12, 20].

In recent work, Papadopoulos *et al.* show that individuals who are blind are better at understanding synthesized speech than individuals who are sighted [18]. Individuals who are blind are also more likely to use synthesized speech as a primary output modality than individuals who are sighted, due to the widespread use of screen readers such as JAWS and WindowEyes. When used as a primary output modality in a screen reader, synthesized speech is often sped up to multiple times real time. However, surprisingly little research has been done on the intelligibility of fast speech for individuals who are blind.

Asakawa and colleagues performed an experiment on the intelligibility of linearly time compressed natural speech in listeners who are blind [2]. They used a recall-based experimental methodology. Their experiment involved only a small number of participants, but they were able to divide their participants into expert users of synthesized speech and novice users of synthesized speech. They found that expert users can recall 90% of SUS content at about 2.5 times, and novice users at about 1.6 times, the default speaking rate of a Japanese TTS system.

Intelligibility Evaluation

[Click here if you need to recap the instructions](#)

1 of 60

Just listen once to this utterance:

Write down what you heard:

(If absolutely unable to guess a word, type xxx for that word.)

Figure 2: SUS screen, which contains a link to the experiment instructions, a button to play the SUS speech, and a text area to type the transcription of the SUS speech

Moos and Trouvain compared fast synthesized speech to linearly time compressed natural speech in sighted listeners and blind listeners who were proficient users of a formant-based synthesizer [16]. In contrast to other research, they found that intelligibility of fast synthesized speech was lower than that of linearly time compressed natural speech for sighted listeners, while higher for blind listeners. However, their method for evaluating intelligibility was via subjective judgments rather than recall, and listeners may not be very good at judging intelligibility. Also, the sighted listeners listened to slower speech than the blind listeners.

Nishimoto and colleagues compared several methods of synthesizing fast speech using a hidden Markov model (HMM) based synthesizer that separately models F0 and phone duration. They compared models trained on fast natural speech and normal-rate natural speech, used in combination with a rapid speech model, which can produce speaking rates from real time to 1.6 times real time. Their experiments included 4 individuals who are blind. They found that models trained on fast natural speech produce more intelligible speech than models trained on normal-rate natural speech as the speaking rate increases [17]. None of these models correspond directly to linearly time compressed speech, so the results are not directly comparable to the other work mentioned here.

In summary, although we have some information about how people who are sighted and people who are blind process fast natural speech and fast synthesized speech, previous research does not give a user or developer of speech synthesis technology guidelines about choosing the “best” synthesis approach or engine, or about where to focus development efforts to get the biggest performance improvement of synthesis systems for individuals with visual disabilities.

3. EXPERIMENT

In this section, we describe the pilot experiment we ran. We followed a proposed standard developed for testing text-to-

Synthesizer	Type	Bandwidth	Voice	Number of participants
CTTS1	concatenative unit-selection	16 bit, mono 16kHz	F	7
			M	6
CTTS2	concatenative unit-selection	16 bit, mono 1600 Hz	F	1
			M	1
FTTS1	formant	16 bit, mono 11025 Hz	F	2
			M	5
FTTS2	formant	16 bit, mono 11025 Hz	F	6
			M	8

Table 1: Synthesizers used, with number of participants per voice per synthesizer

speech intelligibility by the ASA TTS Technology standards committee (S3-WG91).

3.1 Task

We chose a open-response recall task for this experiment. In this type of task, participants listen to a spoken stimulus once and then transcribe what they think they heard. In order to prevent participants from using context and inference to identify words, rather than simply transcribing what they hear, we used **semantically unpredictable sentences** (SUSs) as stimuli. These sentences are grammatically correct but semantically meaningless. Examples include *A polite art jumps beneath the arms* and *The law that finished shows the boots*.

The experiment was web-based. Participants were presented first with a screen containing experimental instructions, second with a form requesting demographic information (see Figure 1), then with 6 training SUSs, and finally with 198 experimental SUSs. Each SUS was presented on its own screen (see Figure 2). The participant clicked on a button linking to an audio file to hear the SUS, and then typed the transcription in the text box, using ‘xxx’ to indicate regions of unintelligible speech.

3.2 Synthesis Engines

We tested two popular approaches to text-to-speech synthesis: formant-based and concatenative unit-selection. A *formant-based synthesizer* generates speech by systematic variation of parameters including formant frequencies and amplitudes, voicing and noise to synthesize audio for each phoneme in the input. Formant synthesizers are widely used in screen readers, and are thought by many in the blind user community to be ‘clearer’ than other synthesizers. Well-known formant synthesizers include eSpeak [6], ETI-Eloquence [9] and DECTalk [8].

A *concatenative unit-selection synthesizer* operates by selecting and concatenating speech units (uniform- or variable-length speech segments) from a speech database to match the input. Generally speaking, concatenative unit-selection synthesizers produce the most natural-sounding speech. However, signal processing is required to speed the output of a concatenative unit-selection synthesizer beyond rates that a human speaker can produce. Well-known concatenative unit-selection synthesizers include AT&T Natural Voices [3] and IVONA [10].

Both approaches to speech synthesis typically incorporate additional processes, e.g. to assign prosody or pronuncia-

tions to input text, and to speed up output speech before play-back. These processes may have substantial impact on the quality of the output speech, so it is hard to isolate the performance of the synthesis algorithm from the performance of the synthesis engine as a whole.

3.3 Participants

Participants were recruited via email and the web. We sent email announcements to numerous organizations including the American Foundation for the Blind and the National Federation of the Blind; some of these organizations graciously forwarded our announcement to their membership.

We restricted participation in this experiment to individuals with early-onset blindness (onset at ≤ 6 years of age), because such individuals are likely to have different hearing abilities from those with late-onset vision impairment [7].

Thirty-six participants completed our pilot study. Four were under 25 years of age, twenty-five were between 25 and 50 years of age, and seven were between 51 and 65 years of age. Four described themselves as using text-to-speech synthesis systems ‘never’ or ‘occasionally’, four as using them ‘frequently’, twenty-five as using them ‘very frequently’, and three were not sure. The most commonly listed text-to-speech engines used were ETI-Eloquence (fourteen participants), JAWS (which comes with ETI-Eloquence by default, nine participants), and Apple VoiceOver’s default voice, Alex (six participants). Thirty-four self-reported as being native speakers of English. Thirteen listened using speakers, while twenty-three used headphones.

3.4 Materials

For this experiment, we used two widely-available formant-based synthesizers and two widely-available concatenative unit-selection synthesizers (see Table 1). Because male and female voices have different characteristics which may affect intelligibility [13], we used male and female American English voices for each synthesizer.

We used the University of Delaware SUSgen system [4] to generate our SUSs. We used 204 SUSs (6 training SUSs, 198 test SUSs) covering the full range of phonemes observed in English. SUS length ranges from 5 to 7 words with a maximum of 10 syllables. Sentence frames include a variety of syntactic structures.

Each synthesizer produced speech for each SUS at six different speeds ranging from 300 to 550 words per minute at intervals of 50 words per minute. This corresponds to

roughly 1.5 times real time to 3 times real time (where ‘real time’ is the synthesizer and voice’s default speaking rate). SUSs were saved as .wav files and played back through the browser in a pop-up window that opened when the participant clicked a link in the SUS screen (see Figure 2). The assignment of test SUSs to speeds, and the order of SUSs, was randomized across participants. The test progressed in blocks of 33 SUSs from 300 words per minute to 550 words per minute in 6 50 word-per-minute steps. Listeners were assigned a text-to-speech system and voice according to a predetermined round-robin order.

3.5 Data Processing

The transcriptions were automatically processed to remove punctuation and replace upper case letters with lower case letters. For the analysis reported in Section 4.1, the transcriptions were also automatically processed to remove typographic errors and unify spelling of homophones, using a word list constructed by hand by one of the authors.

3.6 Issues With Experimental Design

Participants faced several issues with this experimental design. The two biggest issues had to do with audio playback and experiment length. Numerous participants dropped out of our pilot study due to these issues, giving an unbalanced data set.

Participants could only listen to each spoken SUS once. Unfortunately, many screen readers indicate using speech that a link has been clicked on, and this speech would overlap with the SUS speech. We solved this by forcing a 5-second pause between link clicking and start of audio playback; this time is long enough for the screen readers used by our participants to finish speaking.

The second issue was the length of the experiment. It takes about one hour to listen to all 198 test SUSs. This is quite a long time for a web-based experiment; furthermore, due to the cognitive load of listening to and transcribing semantically unpredictable sentences while interacting with a screen reader and various pop-up windows, participants were strongly encouraged to break up the task into multiple sessions. In our current version of this experiment (see Section 5), we have reduced the task length to 60 test SUSs.

4. RESULTS AND DISCUSSION

In this section, we report the results of our pilot study for our main variables of interest (synthesizer type, voice gender and speaking rate). We also report informal post-hoc analyses of participant-related factors, and a comparison of alternative methods for computing transcription accuracy.

4.1 Experiment Results

For our pilot experiment, we measured transcription accuracy using the cosine similarity between a reference transcription for each SUS and the participant’s post-processed transcription for the SUS (see Section 4.3). Figure 3 shows a plot of results by synthesizer type and voice. FTTS2 appears to perform the best, with transcription accuracy over 0.8 for the male voice across all speaking rates. By contrast, FTTS1 shows the steepest rate of decline as speaking rate increases, from above 0.8 at 300 words per minute to below

0.4 at 550 words per minute for both male and female voices. Performance for CTTS1 and CTTS2 also declines as speaking rate increases, though less than for FTTS1. Except for FTTS1, transcription accuracies appear to be higher for the male voice for all synthesizers.

We ran a three-way mixed Anova with two between-subjects variables: synthesizer type (2 levels, formant and concatenative unit-selection) and voice gender (2 levels, M and F), and one within-subjects variable: speaking rate (5 levels, words per minute ranging from 300 to 500 at intervals of 50²). There was a significant main effect for speaking rate, $F(4, 5760) = 59.9759$, $p < .001$, such that transcription accuracy declined as speaking rate increased. There was a trend towards a main effect for synthesizer type, $F(1, 5760) = 3.2563$, $p = .08$. There was no significant main effect for voice gender, and there were no significant interaction effects.

Looking at Figure 3, we see that one of the formant-based synthesizers outperforms the other three, while the other formant-based synthesizer shows the steepest decline in performance as speaking rate increases. The performance of the male voice for one of the concatenative unit-selection synthesizers is similar to that of the female voice for the best-performing formant-based synthesizer. We did post-hoc analyses using t-tests to compare the performance of individual synthesizers, and found significant differences in performance between FTTS1 and FTTS2 ($df = 2143$, $p < .001$), between CTTS2 and FTTS2 ($df = 443$, $p < .001$), and between CTTS1 and FTTS2 ($df = 4190$, $p < .001$).

In general, Figure 3 seems to show that the male voices outperform the female voices except for the poorer-performing formant-based synthesizer. The lack of a significant main effect for voice gender may be due to the unbalanced nature of our data.

Interestingly, although a speaking rate of 500 words per minute is 2.5 times real time, transcription accuracies for all synthesizers were still at or above 50%, in the range Asakawa *et al.* define as acceptable [2].

4.2 Participant-Related Factors

We conducted post-hoc analyses to see if participant-related factors mentioned in the literature also played a role in text-to-speech intelligibility in our data.

Figure 4 shows average transcription accuracy across different speaking rates for different participant age ranges. Participants under 25 years of age had the highest transcription accuracies, and transcription accuracy declined more gradually as speaking rate increased (from 0.89 at 300 words per minute to 0.71 at 550 words per minute). Participants over 51 years of age had the lowest transcription accuracies (from 0.74 at 300 words per minute to 0.37 at 550 words per minute). These results agree with those reported in the literature (e.g. [12]).

²Due an error in the experimental setup, we are missing data for the 550 words per minute setting for CTTS1-M, so we excluded all the 550 words per minute data from this analysis.

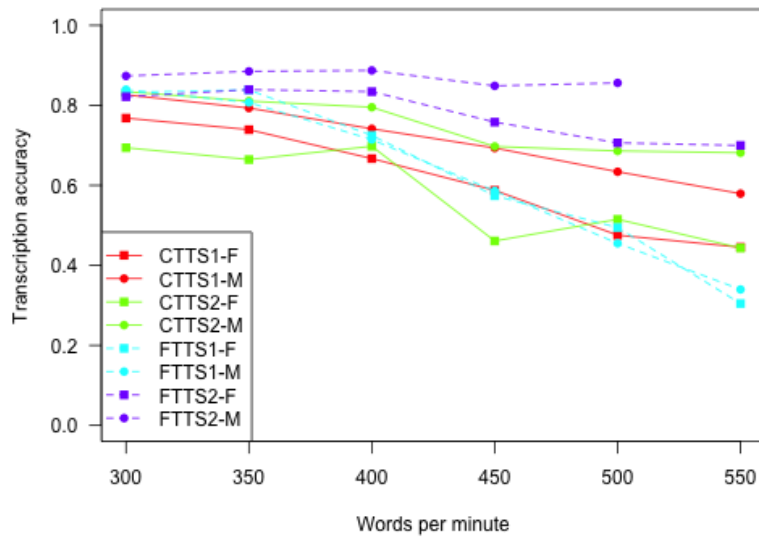


Figure 3: Line graph showing transcription accuracy by synthesizer, voice gender and speaking rate

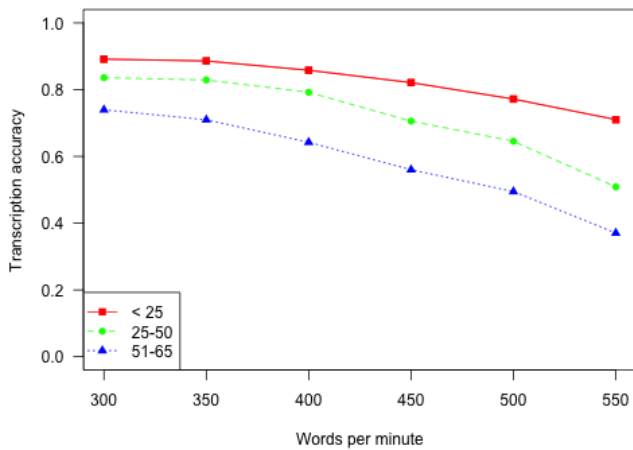


Figure 4: Line graph showing transcription accuracy by speaking rate for different participant age ranges: under 25, 25-50, and 51-64

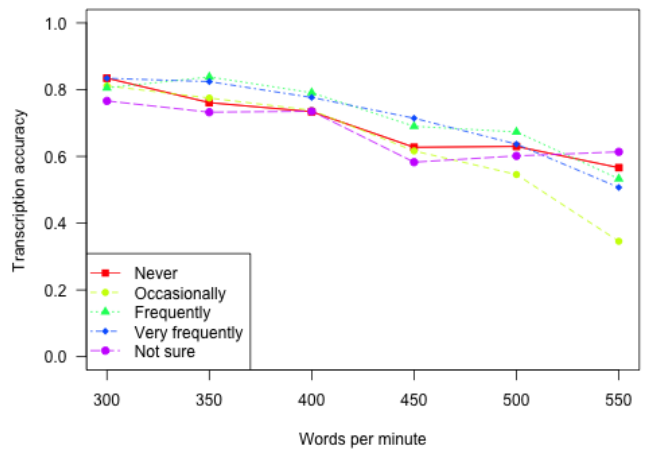


Figure 5: Line graph showing transcription accuracy by speaking rate for expert and non-expert users of text-to-speech synthesizers

Figure 6 shows average transcription accuracy across different speaking rates for expert and non-expert users of text-to-speech synthesizers. Participants who used synthesized speech ‘frequently’ or ‘very frequently’ had the highest transcription accuracies from 350 words per minute (0.83) to 500 words per minute (0.66). Participants who ‘never’ or ‘occasionally’ used synthesized speech, or who were not sure, had lower transcription accuracies from 350 words per minute (0.76) to 500 words per minute (0.59).

Nine of the participants assigned to one of the voices of

FTT2 self-reported as using FTT2. This may partly explain the very good results for FTT2. In fact, familiarity with a synthesizer may negate or severely retard the negative impact on intelligibility of speaking rate; for the FTT2 male voice, transcription accuracy remains above 0.8 even at 500 words per minute.

Figure 6 shows average transcription accuracy across different speaking rates for native and nonnative speakers of English. Native speakers of English achieved higher transcription accuracies at every speaking rate; furthermore, tran-

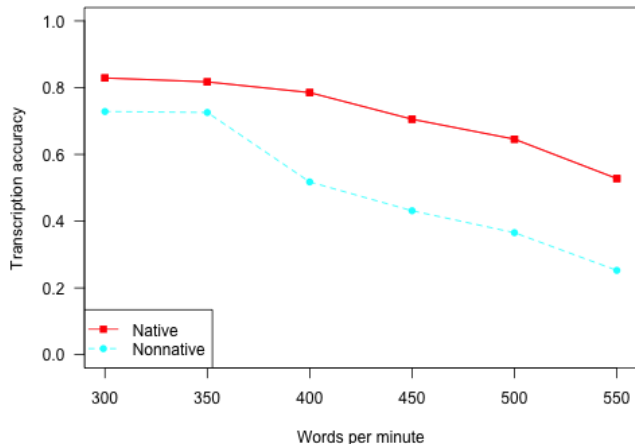


Figure 6: Line graph showing transcription accuracy by speaking rate for native and nonnative speakers of English

scription accuracy declined more slowly as speaking rate increased.

Figure 7 shows average transcription accuracy by listening equipment used by the participant across different speaking rates. The lines are very similar, diverging only at 550 words per minute. The potential impact of equipment appears to be small except at the fastest speaking rates. It may be that participants who use screen readers regularly have good sound equipment.

4.3 Measuring Transcription Accuracy

Transcription accuracy can be measured as orthographic accuracy (ability of the listener to reproduce exactly the words that form the stimulus), corrected orthographic accuracy (ability of the listener to reproduce the words that form the stimulus, post-corrected for typographic errors and for homophones), or phonetic accuracy (ability of the listener to reproduce the sounds in the stimulus). Furthermore, there are multiple different ways to compare a reference transcription (orthographic or phonetic) with a listener’s transcription. In the analyses above, we used corrected orthographic accuracy, measured using cosine similarity. However, in order to do this we had to post-process the transcriptions to account for homophones (e.g. *pear*, *pair*, *pare*) and spelling errors (e.g. *against*, *againest*, *aginst* or *apartment*, *appartment*). Participants in this study frequently made spelling errors that involved dropping a letter or doubling a letter, perhaps because they were listening to a screen reader repeat each character they typed. It takes time for an experimenter to construct a list of permissible substitutions to post-process a set of transcriptions, and even then, word boundary errors (e.g. *about* vs. *a boat*) persist³.

³The ASA TTS Technology standards committee’s proposed standard calls for word boundary errors to be counted as errors.

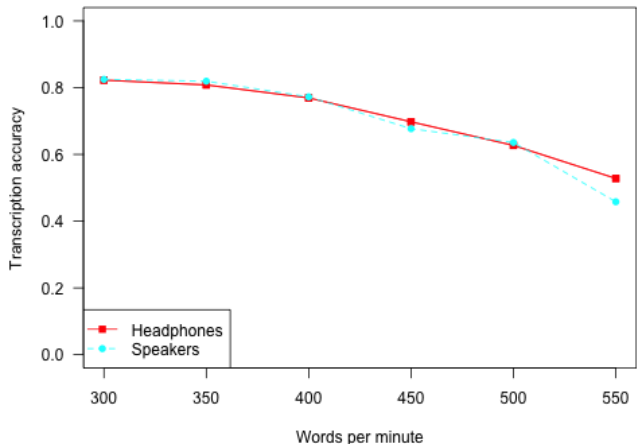


Figure 7: Line graph showing transcription accuracy by listening equipment (headphones or speakers) and by speaking rate

We explored using other measures of orthographic or phonetic accuracy. To compute a quasi-phonetic representation of the transcriptions, we used Double Metaphone [19], as implemented in the Apache commons codec [1]. To compute accuracy, we tried the following metrics, all implemented in the freely-available Simmetrics package [5]:

- **Cosine similarity** – a word-based similarity metric that computes the cosine distance between two vectors representing the input strings
- **Dice** – a character bigram-based similarity metric; given input strings s_1 and s_2 , the dice coefficient is $\frac{2 * |common_bigrams(s_1, s_2)|}{(|bigrams(s_1)| + |bigrams(s_2)|)}$
- **Jaro Winkler** - a character-based similarity metric that takes into account transposition errors; given input strings s_1 and s_2 , the Jaro Winkler score is $\frac{1}{3} * \left(\frac{|matching_chars(s_1, s_2)|}{|chars(s_1)|} + \frac{|matching_chars(s_1, s_2)|}{|chars(s_2)|} + \frac{(|matching_chars(s_1, s_2)| - |transpositions(s_1, s_2)|)}{|matching_chars(s_1, s_2)|} \right)$
- **Levenshtein distance** - a character-based similarity metrics that counts the number of insertions, deletions and substitutions required to turn input string s_2 into input string s_1

In all cases, we used versions of these metrics normalized to the length of the reference transcription.

We automatically processed each participant’s transcription to obtain:

- **Ortho-plain** – a minimally-processed orthographic transcription (remove punctuation and replace upper with lower case)

- **Ortho-processed** – similar to ortho-plain, except that a hand-created list of permissible substitutions covering common typographic errors and homophones was also applied
- **Phono-plain** – a double-metaphone representation of ortho-plain

Since Double Metaphone produces a single word as output, and represents the order of phonemes in the transcription, it does not make sense to use any metrics other than Levenshtein on the phono-plain version of a transcription. In addition, since we hope to replace a manually-constructed list of substitutions with a robust similarity metric, it only makes sense to compute Cosine on the ortho-processed version of a transcription. Consequently, we computed similarities as shown in Table 2.

The **Cosine** metric represents our best human-aided effort at a similarity metric, so we are looking for an alternative metric that is very closely correlated with **Cosine**. Table 3 shows the Pearson correlations between **Cosine** and the other similarity computations summarized in Table 2⁴. The good news is that both **Cosine-plain** and **Dice-plain** are highly correlated with **Cosine**; this means that we probably do not need to hand-author lists of spelling corrections and homophone substitutions. **JaroWinkler-plain** has the lowest correlation with **Cosine**: participants in this experiment tended to drop or repeat letters, not swap them. **Levenshtein-plain** is highly correlated with **Cosine**, but not as highly as **Cosine-plain** or **Dice-plain**.

Levenshtein-phono is also highly correlated with **Cosine**, though not as highly as we might expect. An examination of the data indicates that the rules Double Metaphone uses to create a quasi-phonetic representation of a string are over-generous for our purpose, for example mapping word-final /d/ and /t/ to the same character (while we want, e.g. *send* and *sent* to be treated as different words). This means that the phono-plain representations of two quite different strings will be quite similar: consider *why should the velmas in the mont* vs. *why should the bell miss an amount* (**Cosine**: .463; **Levenshtein-phono**: .857), or *the bags left between a small coat* vs. *the bags leapt beneath the small coat* (**Cosine**: .617; **Levenshtein-phono**: .75). In future work, we may consider more direct phonetic representations such as a phonetic transcription from the input to a text-to-speech synthesizer not included in our experiments.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we report the results of a pilot study examining the impact of synthesis method, voice gender and speaking rate on intelligibility of fast synthesized speech for individuals with early-onset blindness. We found a significant effect of speaking rate and a trend towards significance for synthesizer type. In post-hoc analyses, we found that participant-related factors also affect intelligibility of fast synthesized speech. In particular, we confirmed prior research showing that age affects ability to understand fast

⁴The order of similarity of the other metrics to **Cosine** remains the same when Spearman rank correlations are used as the measure of similarity.

speech. We also found some evidence that familiarity with a synthesizer and voice may ameliorate the negative impact of speaking rate on intelligibility of synthesized speech.

We are currently conducting an experiment that uses a modified version of the methodology used in our pilot study. In particular, we have corrected issues pertaining to interference between a screen reader and stimulus playback, and have shortened the study. We are also including text-to-speech engines that use a third synthesis method: Hidden Markov Model (HMM)-based synthesis [21]. We intend to collect data for this experiment until we have a balanced data set covering two voices (male and female) for each of at least two synthesizers for each of these three approaches to text-to-speech synthesis. The ASA TTS Technology standards group is conducting a parallel study (using the same experimental methodology) with participants who do not have early-onset blindness.

In future work, we would like to further separate the synthesis method from other components of the text-to-speech synthesizer, particularly the signal processing for speeding up speech. We can do this by collecting speech at a single rate from multiple synthesizers (and, optionally, a human speaker) and speeding it up as a separate process using different methods, including linear speed up and linear speed up excluding pauses [16].

In this experiment, we evaluated intelligibility using a recall task, which is an approximation of close reading of a text. People with visual disabilities who use text-to-speech synthesis as a primary output modality also use the speech stream for skimming/ gisting (e.g. skimming through emails, skimming through a news story) and for search. It would be interesting to look at whether the synthesis method or method of speeding up speech affect usability of synthesized speech for these and other alternative use cases. For example, a study could be run in which participants are asked to listen for a particular phrase or topic and then stop the playback as soon as possible. Perhaps skimming and searching are possible at higher speaking rates than close reading. We leave these questions for future research.

6. ACKNOWLEDGMENTS

We thank the members of the ASA TTS Technology standards committee (S3-WG91) for providing their proposed standard method for evaluating TTS intelligibility, and for letting us borrow their materials for our study. We thank the individuals who contributed text-to-speech synthesis output for our experimental materials. We also gratefully acknowledge the experimental participants, who not only worked patiently through our long experiment but also sent us detailed feedback on the interactions between the experiment pages and their screen readers.

7. REFERENCES

- [1] Apache commons codec.
<http://commons.apache.org/codec/apidocs/overview-summary.html>.
- [2] C. Asawka, H. Takagi, S. Ino, and T. Ifukube. Maximum listening speeds for the blind. In *Proceedings of the International Conference on Auditory Display*, 2003.

Metric	Transcription Type		
	Ortho-plain	Ortho-processed	Phono-plain
Cosine	Cosine-plain	Cosine	
Dice	Dice-plain		
Jaro Winkler	JaroWinkler-plain		
Levenshtein	Levenshtein-plain		Levenshtein-phono

Table 2: Similarity computations made

	Cosine-plain	Dice-plain	JaroWinkler-plain	Levenshtein-plain	Levenshtein-phono
Cosine	.982	.982	.699	.911	.897

Table 3: Correlation between Cosine and other similarity computations summarized in Table 2

- [3] M. Beutnagel et al. The AT&T next-gen TTS system. In *Proceedings of the Joint Meeting of the ASA, EAA and DAGA*, 1999.
- [4] H. T. Bunell and J. Lilley. Analysis methods for assessing TTS intelligibility. *Presented at the 6th ISCA Workshop on Speech Synthesis*, 2007.
- [5] S. Chapman. Simmetrics. <http://staffwww.dcs.shef.ac.uk/people/S.Chapman/simmetrics.html>.
- [6] eSpeak. <http://espeak.sourceforge.net/>.
- [7] F. Gougoux et al. Neuropsychology: Pitch discrimination in the early blind. *Nature*, 430, 2004.
- [8] W. Hallahan. DECTalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, 7(4), 1995.
- [9] S. Hertz, R. Younes, and N. Zinovieva. Language-universal and language-specific components in the multi-language eti-eloquence text-to-speech system. In *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.
- [10] IVONA. <http://www.ivona.com/>.
- [11] E. Janse. Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42:155–173, 2004.
- [12] E. Janse, M. van der Werff, and H. Quené. Listening to fast speech: aging and sentence context. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007.
- [13] D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.
- [14] J. Lebeter and S. Saunders. The effects of time compression on the comprehension of natural and synthetic speech. *Working Papers of the Linguistics Circle of the University of Victoria*, 20:63–81, 2010.
- [15] D. Moers, P. Wagner, B. Möbius, F. Müllers, and I. Jauk. Integrating a fast speech corpus in unit selection synthesis: experiments on perception, segmentation, and duration prediction. In *Proceedings of Speech Prosody*, 2010.
- [16] A. Moos and J. Trouvain. Comprehension of ultra-fast speech - blind vs. “normally hearing” persons. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007.
- [17] T. Nishimoto et al. Effect of learning on listening to ultra-fast synthesized speech. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference*, 2006.
- [18] K. Papadopoulos, E. Katemidou, A. Koutsoklenis, and E. Mouratidou. Differences among sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. *Augmentative and Alternative Communication*, 26(4):278–288, 2010.
- [19] L. Phillips. The double metaphone search algorithm. *C/C++ Users Journal*, 18(6):38–43, June 2000.
- [20] B. Sutton, J. King, K. Hux, and D. Beukelman. Younger and older adults’ rate performance when listening to synthetic speech. *Augmentative and Alternative Communication*, 11(3):147–153, 1995.
- [21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *Proceedings of ICASSP*, 2000.