INCORPORATING SYNTHETIC SPEECH INTO A PHONEME-BASED COMMUNICATION SYSTEM

Ha Trinh, Annalu Waller, Rolf Black
School of Computing, University of Dundee

ABSTRACT

This study investigated the potential of integrating synthetic speech into a phoneme-based communication aid, which enables nonspeaking individuals to access 42 phonemes and blend phonemes into spoken words. Evaluations of two synthetic voices produced by a phoneme-to-speech synthesizer implemented within the study showed that both voices achieved relatively high degrees of speech intelligibility.

BACKGROUND

The importance of voice output communication aids (VOCAs) in improving quality of life for individuals with severe speech and physical impairments (SSPI) have been well documented in the literature [2]. Most of the commercially available VOCAs can be classified into two categories, namely pictographic-based and letter-based systems. Pictographic-based VOCAs employ graphical symbol systems to encode a limited vocabulary of commonly used words and phrases, allowing users to produce spoken output through the selection of symbol sequences. Thus, users are restricted to retrieve pre-stored items rather than being able to create novel words spontaneously. To overcome this limitation, a number of letter-based VOCAs have been developed to enable users to spell their own messages and generate them in the form of synthesized speech. However, this type of VOCA requires users to master literacy skills, making it unusable for a large proportion of nonspeaking people who experience literacy difficulties [4].

Our research aims to address the abovementioned issues of existing VOCAs by exploring the potential of developing a phoneme-based communication system for individuals with SSPI. The proposed system allows users to access the 42 phonemes introduced in the Jolly Phonics literacy learning program [5] and combine phonemes into spoken words and messages [1]. These phonemes have been selected to incorporate into the system due to their letter-sound correspondence. Since the system uses spoken phonemes (i.e. English sounds) as the base units for speech generation, it provides users with an unlimited vocabulary, enabling them to create spontaneous, novel words and messages without being literate. Therefore, it has the potential of providing effective communication support for nonspeaking individuals with limited or no literacy skills in interactive conversation. Moreover, as the system employs a set of phonemes that have been widely used for literacy teaching throughout the UK, it can also be utilized as an educational tool to assist preliterate children with SSPI in literacy learning.

One of the most essential requirements for the development of such a phoneme-based communication system is to identify an efficient method of generating intelligible speech output from phoneme input. The need for being able to produce speech from any phoneme sequences suggests that digitized speech is inappropriate for the system as it would be impractical to pre-record and store all possible combinations of the 42 phonemes. Therefore, the present study aims to investigate whether synthetic speech can be incorporated into the system to allow for automatic speech generation from phoneme sequences. Although text-to-speech synthesizers (TTS) have been extensively used in many existing VOCAs [2], there have been no published studies to date which examine how well these TTS could perform with a restricted set of 42 phonemes that are specifically intended for literacy teaching.

RESEACH OBJECTIVES

The present study addressed the research question of whether state of the art speech synthesizers can be integrated into a phoneme-based communication system to produce highly intelligible speech, given input selected from the set of 42 phonemes used in the Jolly Phonics literacy learning program [5]. To answer this question, the study investigated the feasibility of accessing and customizing existing text-to-speech synthesis systems at phonemic level to implement a phoneme-to-speech synthesizer. Speech perception tests were conducted to evaluate the intelligibility of synthetic speech produced by the resulting synthesizer in comparison with natural speech.

IMPLEMENTATION OF A PHONEME-TO-SPEECH SYNTHSIZER

Figure 1 shows the architecture of the phoneme-to-speech synthesizer implemented within the study. Two text-to-speech (TTS) engines were included in the synthesizer, including the CereVoice®[1] and Microsoft TTS[2]. These are two high quality unit-selection TTS engines that have currently been integrated into a number of commercial applications. Microsoft Speech API (SAPI5) was used to access the two TTS engines at phonemic level.

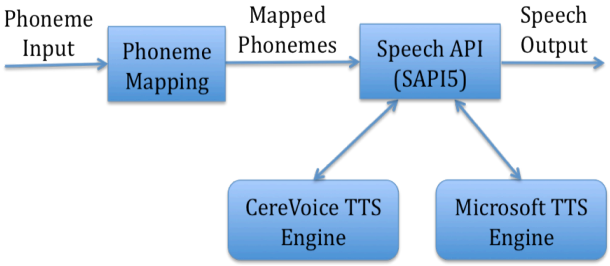

Figure 1: Overall architecture of the phoneme-to-speech synthesizer

The synthesizer consists of a Phoneme Mapping component, which converts the Jolly Phonics's phonemes used for educational purposes to the corresponding phonemes used in TTS engines. Due to the variation in phoneme sets across different TTS engines, it was essential to implement a separate phoneme mapping module for each engine incorporated in the system. Output of the phoneme mapping procedure would then be streamed into the digital signal processing (DSP) component of either the CereVoice or Microsoft TTS engine via the SAPI5 interface to generate acoustic waveforms of the synthesized speech.

EVALUATION METHODOLOGY

A speech perception study was carried out to compare the speech intelligibility of three voices, including: (1) synthetic voice produced by the phoneme-to-speech synthesizer using the CereVoice speech synthesis engine; (2) synthetic voice produced by the phoneme-to-speech synthesizer using the Microsoft TTS engine; (3) human voice. The study was focused on the phoneme and single word intelligibility, which was determined by the ability of human listeners to correctly identify isolated spoken words.

Subjects

Thirty English native speakers aged from 18 to 60 years old with reported normal hearing were recruited from within a university to participate in the study.

Evaluation procedure

The evaluation consists of two sessions, as described below:

Session 1: Closed-Response Modified Rhyme Test

- *Method*

In this session participants were asked to complete a closed-response Modified Rhyme Test (MRT), a well-established test for evaluating the phoneme intelligibility of synthetic speech [3]. Each participant was asked to listen to a list of 50 monosyllabic words, each of which consists of an initial consonant, a middle vowel, and a final consonant. After hearing each word, they were required to select the word they think they heard from a set of six similar-sounding words, which differ from the correct answer by a single phoneme in either initial or final position.

[1] CereProc Ltd., Edinburgh, EH8 9LE, UK
[2] Microsoft Corp., Redmond, WA 98052-7329, USA

- *Stimuli*

50 monosyllabic words in list A of the MRT 300-word set were selected for testing. The selected words were recorded in three voices, including two synthetic voices (i.e. CereVoice and Microsoft TTS) and a female English native speaker. The recording of the synthetic voices were completed using a function of the phoneme-to-speech synthesizer, which produces synthesized output from phoneme sequences and saves the output to .wav files. The recording of the natural speaker was setup in a quiet room where the speaker was seated in front of a stereo microphone. The speaker was asked to wear a headphone and read the 50 words continuously into the microphone, which was plugged into a computer running an audio recording software tool. The recorded speech was then segmented into 50 .wav files corresponding to the 50 words in the word list.

- *Procedure:*

Three groups of ten participants each were randomly assigned to three conditions: CereVoice, Microsoft TTS, and human voice. The experiment was conducted in a quiet environment where the participants were seated in front of a computer which runs the MRT program, wearing headphones. On-screen instructions explaining how to complete the MRT program were provided. The participants were also given time to practice with the program prior to the actual test.

Session 2: Open-response Word Recognition Test (WRT)

- *Method*

This test aimed to evaluate the intelligibility of the synthetic speech at word level. The participants were asked to listen to a list of 90 words of varying phonemic lengths and complexity, including both mono- and multi-syllabic words. After hearing each word, the participants were instructed to enter the word they think they heard into a textbox instead of choosing the correct answer from a closed-response set.

- *Stimuli*

Thirty 90-word lists were prepared, one for each participant. Each word list was a combination of three 30-word sub-lists, each of which was spoken by one of the three tested voices (i.e. CereVoice, Microsoft TTS, and human voice). These 30-word sub-lists were randomized from a 150-word base list, which was selected from the Schonell's Essential Spelling list [6] and recorded with the three tested voices using the same methods and equipments described in session 1. All the randomized word sub-lists were automatically checked using a computer program to ensure that they are compatible in terms of phonemic lengths and complexity. The word ordering in each 90-word list was also randomized to ensure that there were no more than two consecutive words spoken by the same voice.

- *Procedure*

Upon the completion of the first session, all 30 participants were asked to take part in session 2, which was conducted using the same procedure as in session 1 with on-screen instructions and practice sessions followed by the actual test. At the end of the session the participants were invited to check the results with the researchers to verify whether their incorrect answers were results of misspelling or mishearing.

RESULTS AND DISCUSSION

The speech intelligibility of the tested voices was evaluated in two dimensions, including phoneme intelligibility and single word intelligibility. The phoneme intelligibility was determined by the mean percentage of correct answers for the closed-response Modified Rhyme Test (MRT). The single word intelligibility was determined by the mean percentage of correct answers for the open-response Word Recognition Test (WRT).

Overall, the results demonstrated a relatively high degree of both phoneme and single word intelligibility of the two synthetic voices produced by the phoneme-to-speech synthesizer. The CereVoice voice scored a mean percent correct of 92.20% in the MRT and 84.06% in the WRT, whilst the Microsoft TTS voice obtained the percent correct of 87.8% and 86.44% for the MRT and WRT, respectively.

Two one-way ANOVA tests were performed to analyze the differences in speech

intelligibility among the three tested voices. Results showed that both phoneme and single word intelligibility differed significantly across the three voices ($F(2, 27) = 30.34$, $p < .001$ for the MRT and $F(2, 87) = 33.40$, $p < .001$ for the WRT). Post-hoc analyses using Tukey tests indicated that the human voice was significantly more intelligible than the two synthetic voices at both phoneme and single word levels ($M = 98.80$, $SD = 1.40$ for the MRT, and $M = 97.89$, $SD = 3.09$ for the WRT, $p < .05$). The phoneme intelligibility of the CereVoice voice ($M = 92.20$, $SD = 2.74$) was significantly higher than that of the Microsoft TTS voice ($M = 87.80$, $SD = 4.56$), $p = .012$. However, the difference between the single word intelligibility of these two synthetic voices was not statistically significant at $p < .05$.

Further analysis was conducted on the results of the MRT to compute recognition error rates for different phoneme groups, which could help identify phoneme error patterns for the tested synthetic voices. Table 1 presents data on the percent errors by phoneme groups for the two synthetic voices (note that vowels were not included in the analysis as the MRT was focused on testing the intelligibility of consonants which were considered more problematic for speech synthesizers [3]). High percent errors on stop sounds (/b/, /g/, /d/, /p/, /t/, /ck/) were reported for both CereVoice and Microsoft TTS voices. CereVoice also had a great number of errors on nasal sounds (/n/, /m/, /ng/), whilst both voices achieved very low error rates for approximants (/w/, /r/, /l/). These results were consistent with the findings reported in previous studies on phoneme intelligibility of synthetic voices [3, 7].

Table 1: Percentage of recognition errors by phoneme groups for two synthetic voices

Phoneme Group	**Percent Error (%)**	
	CereVoice	***Microsoft TTS***
Stops	13.33	26.39
Fricatives and affricates	5.0	4.51
Approximants	0	0.83
Nasals	20	5.56

CONCLUSION

Results from an evaluation of two synthetic voices produced from a phoneme-to-speech synthesizer developed within the study indicated that these voices have obtained relatively high degrees of speech intelligibility. This demonstrated the potential of incorporating synthetic speech into phoneme-based communication systems. However, it was shown that there was still a considerable gap between the intelligibility of phoneme-based synthetic speech and that of natural speech. Phoneme error analyses revealed that synthetic voices tend to perform poorly on stop and nasal sounds. Thus, further work is needed to improve the modeling of these sounds in the speech synthesizer.

ACKNOWLEDGEMENTS

The authors would like to thank SICSA (Scottish Informatics and Computer Science Alliance) and the School of Computing, University of Dundee for funding the research. We would also like to thank CereProc Ltd. for providing us with a free copy of their SAPI5-compliant CereVoice® TTS engine.

REFERENCES

[1] Black, R., Waller, A., Pullin, G., Abel, E., Introducing the PhonicStick: Preliminary evaluation with seven children. in *13th Biennial Conference of the International Society for Augmentative and Alternative Communication* (Montreal, Canada, 2008).

[2] Glennen, S.L., DeCoste, D. C. *The Handbook of Augmentative and Alternative Communication*. Thomson Delmar Learning, 1997.

[3] House, A., Williams, C., Hecker, M., Kryter, K. Articulation testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, *37*. 158-166.

[4] Koppenhaver, D.A., & Yoder, D. E. Literacy issues in persons with severe speech and physical impairments. in Gaylord-Ross, R., Ed. ed. *Issues and research in special education*, Teachers College Press, Columbia University, New York, NY, 1992, 156-201.

[5] Lloyd, S.M. *The Phonics Handbook*. Jolly Learning Ltd., Chigwell, 1998.

[6] Schonell, F.J. *The essential spelling list*. Nelson Thornes, 2000.

[7] Venkatagiri, H. Phoneme Intelligibility of Four Text-to-Speech Products to Nonnative Speakers of English in Noise. *International Journal of Speech Technology*, *8* (4). 313-321.