Synthesizing Speech using the AusTalk Corpus

Zhijie Shao, Richard E. Leibbrandt, and Trent W. Lewis

The Centre for Knowledge and Interaction Technologies
School of Computer Science, Engineering, and Mathematics,
Flinders University, Adelaide, Australia

shao0030@flinders.edu.au, richard.leibbrandt@flinders.edu.au, trent.lewis@flinders.edu.au

Index Terms: speech synthesis, MARY TTS, Blizzard Challenge, AusTalk

1. Introduction

Speech Synthesis is also called text-to-speech (TTS) and its main task is to produce speech (acoustic waveform) from text.[1] It has been widely used in various domains. For example, embodied conversational agents adopt the speech synthesizer to speak and audio books can “read” text for the visually-impaired. We report on the first attempt to use the AusTalk corpus [4] to synthesize an Australian accented voice – few, if any exist in the public domain. This study adopted MARY (Modular Architecture for Research on speech sYnthesis)[2], a speech synthesis system that includes useful auxiliary functions such as Voice Import Tool and Emotion Markup Language Support. We made use of the Blizzard[3] and AusTalk[4] corpora as sources of data for synthesizing the voices using the Voice Import Tool. We analyze the reasons for the relatively poor quality of the AusTalk voice and propose relevant improvement strategies. The objective of the study is to create a satisfactory “Aussie” voice using resources in hand.

2. Methodology

There are two main ways to create a synthesized voice, namely Unit Selection Synthesis and Hidden Markov Model (HMM) Synthesis.[5] Firstly, Blizzard voice of Unit Selection and Hidden Markov Model were created by utilizing the entire Blizzard corpus consisting of 5884 sentences. Before creating the AusTalk voice, 59 sentences that were recorded in Step 7 of Session 2 in the standard AusTalk recording protocol [4] were adopted as training source. In order to obtain reasonable comparison results, 59 sentences were randomly selected from the Blizzard data to create the voice, although the Blizzard corpus contained a greater quantity of data than AusTalk.

3. Results

The results show that training on all of the Blizzard data can produce a high-quality voice. The experiment tested almost all situations such as number, abbreviation and so on. Furthermore, according to different audio effects, the Blizzard voice also performs well. However, the AusTalk voice performs unsatisfyingly: some phonemes are unclearly pronounced and are sometimes mixed with noise throughout the background. Despite changing the audio effects and tuning the parameters during using Voice Import Tool, the effect of the voice was still unsatisfactory. On the other hand, the voice being created by the subset of Blizzard data performs better than the AusTalk one in most cases. Although it cannot sound exactly same as the voice being generated by the whole Blizzard corpus, the text can be approximately conveyed to the desirable speech.

4. Discussion

One reason why the Blizzard Data can deliver high-quality sound in the synthesized voice appears to be that the quality of the original training audio files is better than that of the AusTalk ones. Through comparing between sources, it is assumed that the Blizzard Challenge recording equipment and environment is much better than the AusTalk recordings. In addition, in terms speaker in the Blizzard Challenge corpus was a native speaker of US English, professional voice talent, voice coach, and singer [3]. The participants of the AusTalk collection were volunteers with no specific voice training. Therefore, clarity and quality of voices was variable [4].

This study only examined a small component of the AusTalk corpus, but there is over three hours of recordings per subject. Further work, could investigate the utility of other components (e.g. digits, single words) to expand the phonetic coverage.

Another reason relates to the accuracy of sound alignment. It is possible for Voice Import Tool not to accurately align phonemes with specific characters in the text, especially with small corpus. Hence, improving the alignment algorithm, or even using hand-aligned data, in MARY should also improve the quality of the synthesized sound.

Preliminary, subjective evaluation by the researchers has suggested that in some cases the sound quality of the AusTalk based voice was comparable to the larger Blizzard based voice. A more systematic user evaluation is planned for the future.

5. Reference

[1] Dutoit, T., *An introduction to text-to-speech synthesis*. Vol. 3. 1997: Springer.

[2] Schröder, M. and J. Trouvain, *The German text-to-speech synthesis system MARY: A tool for research, development and teaching.* International Journal of Speech Technology, 2003. **6**(4): p. 365-377.

[3] Simon, K. and K. Vasilis, *The Blizzard Challenge 2010*. 2010: Blizzard Challenge Workshop.

[4] Alghowinem, S., M. Wagner, and R. Goecke. *AusTalk—The Australian speech database: Design framework, recording experience and localisation*. in *Information Technology in Asia (CITA), 2013 8th International Conference on*. 2013. IEEE.

[5] Jurafsky, D. and H. James, *Speech and language processing an introduction to natural language processing, computational linguistics, and speech.* 2000.