Robot Parrot

Caleb Kaiji Lu
caleb.lu@sv.cmu.edu

Tyler Nuanes
tyler.nuanes@sv.cmu.edu

Serhan Oztekin
serhan.oztekin@sv.cmu.edu

Nanshu Wang
nanshu.wang@sv.cmu.edu

1 Idea Summary

Our project, *Robot Parrot*, aims to implement speech-to-speech conversion for only one subject, X. Specifically, we want any arbitrary sentence to be produced in X's voice. Generalized to reproducing any voice from any input voice, such a technology has a range of implications, from voice dubbing in the entertainment industry to personalizing a computer interface. We will generate the original voice using a Text-to-Speech package that generates a machine-voice.

According to the literature, speech-to-speech conversion has been an active research area for the past 40 years, with current methods often relying on DNN [3] or MLE [11]. According to Mohammadi's paper [9], voice conversion "systems still exhibit deficiencies in accurately mimicking a target speaker spectrally and prosodically, and simultaneously maintaining high speech quality." As such, given the short nature of this project, we do not expect to implement a full solution. Instead we are narrowing down our goal to the specific task of converting a particular machine voice to a single human voice.

2 Literature Review

Voice conversion modifies an audio with source speaker voice to a target speaker voice without changing the content. Currently, a company known as *Lyrebird* [1] is producing software capable of translating machine voices to any human voice at a reasonable quality after a sample of one minute of speaking. It can also be applied in voice conversion between multiple languages [2].

Many papers have been written on this topic [10][9]. One approach of voice conversion is to learn the source—target relationship from a number of utterances [10]. There are parametric methods [7] which model a mapping function from a source to target in some feature spaces. An alternative approach is called unit selection[4][6]. The idea is to select the segments from training sets of a target voice, which correspond to the speech content of source voice, then concatenating the segments with smooth transition.

The training process can be either text-dependent or text-independent[4]. Text-dependent require audio of same sentences of both source and target, then use Dynamic Time Warping to align the recordings and get the mapped acoustic features. Text-independent does not require recordings of

the same sentences, or even the same language. Audios are segmented into frames and clustered into groups of similar features. And the acoustic features of source and target are mapped with same categories. The mapping function can be found in different ways: signal processing, linear algebra, machine learning(both discriminative model, e.g., deep learning, and generative model, e.g., GMM)[8]. We intend to perform text-dependent speech conversion.

Based on the breadth of current and past research, it appears there is no established solution to this problem in either academia nor industry. *Lyrebird* comes close, but it is often still possible to distinguish its audio from the human speaker. As such, we hope to produce reasonable results, but we do not expect speech conversion to be perfect.

3 Project Proposal

Our project goal is to perform speech-to-speech conversion on a machine voice to a human voice. Our project consists of several stages:

1. Implement a text-to-speech package to generate machine voice waveforms.
2. Record a particular speaker X saying a number of sentences.
3. Identify features in speech that carry word information as well as carry the unique "voice" of the speaker.
4. Use machine learning to train on the recordings.
5. Generate new sentences with the text-to-speech package.
6. Use the machine to "reproduce" the speech in the voice of X.

The bulk of the project time will likely be spent in identifying important features as well as in training and testing the machine. Based on our literature review, possible acoustic features could be fundamental frequency, formants, MFCC, and etc.

While evaluation of this project may be quite difficult, we think that it would be impressive to produce a series of sentences through speech-to-speech and compare those to sentences spoken by X. To do so, we will use subjective evaluation, where a number of subjects will listen to each recordings and using 5 value grading of quality of converted recordings and similarity to the real recordings. We will use a standard test proposed by TC-STAR[5] projects using MOS (Mean of Score) as a measure of both quality and similarity. We expected to get above 3.0 on quality and 2.5 on similarity MOS.

References

[1] Lyrebird - create a digital copy of voice.

[2] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. Intent transfer in speech-to-speech machine translation. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 153–158. IEEE, 2012.

[3] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad. Voice conversion using artificial neural networks. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3893–3896. IEEE, 2009.

[4] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno. Voice conversion of non-aligned data using unit selection. *TC-STAR WSST*, 2006.

[5] D. Erro and A. Moreno. Weighted frequency warping for voice conversion. In *Interspeech*, pages 1965–1968, 2007.

[6] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore. Cute: A concatenative method for voice conversion using exemplar-based unit selection. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5660–5664. IEEE, 2016.

[7] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1303–1306. IEEE, 1997.

[8] A. F. Machado and M. Queiroz. Voice conversion: A critical survey. *Proc. Sound and Music Computing (SMC)*, pages 1–8, 2010.

[9] S. H. Mohammadi and A. Kain. An overview of voice conversion systems. *Speech Communication*, 2017.

[10] Y. Stylianou. Voice transformation: a survey. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3585–3588. IEEE, 2009.

[11] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.