English
Research Express@NCKU Volume 27 Issue 6 - September 26, 2014
[http://research.ncku.edu.tw/re/articles/e/20140926/4.html]

Synthesis of Spontaneous Speech With Syllable Contraction Using State-Based Context-Dependent Voice Transformation

Chung-Hsien Wu*, Yi-Chin Huang, Chung-Han Lee, Jun-Cheng Guo

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan

chunghsienwu@gmail.com

IEEE/ACM Trans. Audio, Speech, and Language Processing, DOI (identifier) 10.1109/TASLP.2013.2297018, Vol. 22, No. 3, March 2014, pp. 585~595.

Pronunciation variation plays an important role in spontaneous speech. In this research, the transformation function is adopted for the pronunciation variation modeling, and articulatory features are also considered to predict pronunciation variation. Transformation function is used to generate variation phones and to solve the problem of lacking phone models in previous work. The problem for insufficient number of training data is solved by clustering acoustic variations using articulatory features. In this work, we achieve the goal of synthesizing spontaneous speech by generating pronunciation variations.

This research aims to: 1) importing transformation function into Hidden Markov Model (HMM) to model pronunciation variations, and 2) predicting pronunciation using Decision Tree (DT).

Research Method

This approach proposed a systematic way to employs the Hidden Markov model (HMM), a multi-dimensional linear regression model, as the transformation function to model the relationship between read speech and the corresponding spontaneous speech with syllable contraction. With insufficient number of training data, the obtained transformation functions are categorized using a decision tree (DT) based on linguistic and articulatory features for better and efficient selection of suitable transformation function. Finally, the acoustic features of the synthesized frame sequence are transformed based on the transformation functions and the spontaneous speech is synthesized from the transformed features using the corresponding Mel log spectrum approximation (MLSA) filter.

Speech Pair Aliment

In general, spontaneous speech is uttered in a faster speaking rate compared to the read speech. To elucidate the acoustic features of syllable contraction in spontaneous speech, this study adopts the dynamic time warping (DTW) algorithm to align the read speech with the spontaneous speech with syllable contraction. The Euclidean distance was adopted as the spectral distance measure in the DTW procedure. Fig. 2 shows the results of the alignment using DTW between read speech of the word "這樣/tʂɤ4iɑŋ4/" and the corresponding spontaneous speech with syllable contraction "降/tɕiɑŋ4/". The red line indicates the DTW alignment path. The oblique line (part 1 of Fig. 1) shows the relationship mapping from read speech to spontaneous speech

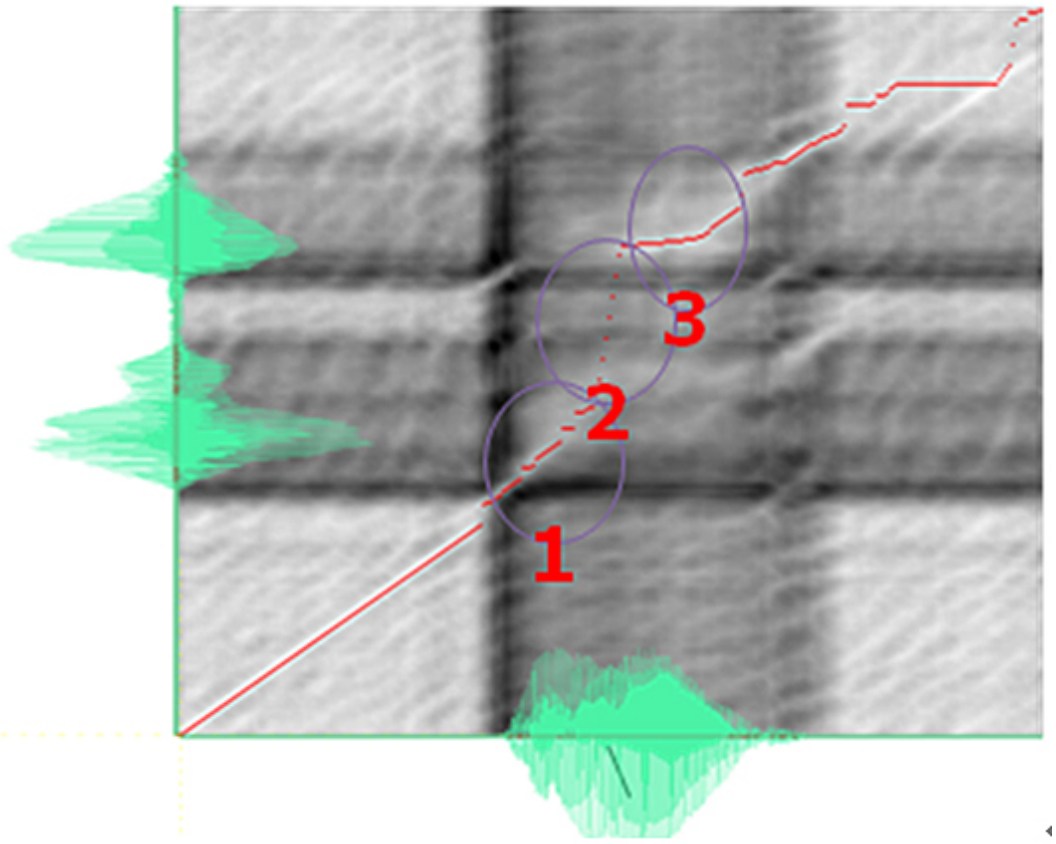

Figure 1 The alignment result using DTW

Transformation function proposed in this study is used to transform the speech sequence of read speech to the corresponding spontaneous speech sequence with syllable contraction. A multi-dimensional linear regression model is adopted as the transformation function to convert the feature vector sequence ***X*** of read speech to the feature sequence ***Y*** of spontaneous speech with syllable contraction. The aligned feature vector sequences from the source (read) speech $X = x^1, x^2,\ldots, x^n$ and $Y = y^1, y^2,\ldots, y^n$ from the target speech (spontaneous speech with syllable contraction) are used as the parallel training data for transformation function construction. . This approach adopts a linear transform to depict the dependency between the two feature vectors (Source *x* and Target *y*):

$$\mathbf{Y} = \mathbf{AX} + \mathbf{R} \tag{1}$$

where **A** is the linear transformation matrix. The residual vector **R** represents the difference between the transformed feature vector, **AX**, using the transformation matrix **A** and the target feature vector ***Y***. The training phase of the transformation functions employs an HMM , which is estimated by maximizing the likelihood function of the joint distribution $P(\mathbf{X}, \mathbf{Y}|\lambda)$ to model the acoustic features of read speech and the corresponding spontaneous speech with syllable contraction, as Fig. 3 shown.

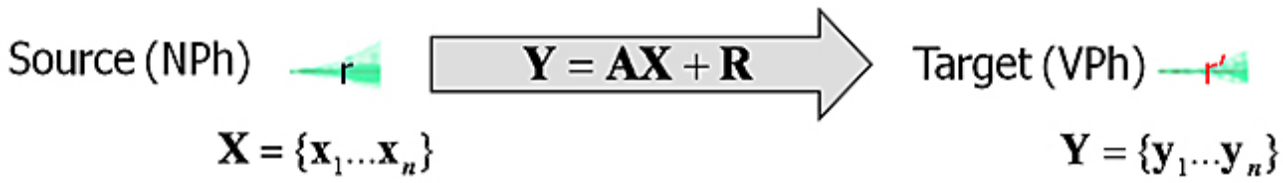

Figure 2 The illustration of the linear transformation function

Fig. 3 shows the operation of the state-based duration model, which can be constructed and integrated into the conventional HMM-based TTS system. The ratio between source duration $\mathbf{L}^{\mathbf{x}}$ and target duration $\mathbf{L}^{\mathbf{y}}$ can be used to scale the duration information. The duration of read speech can be linearly scaled to the duration of spontaneous speech with syllable contraction.

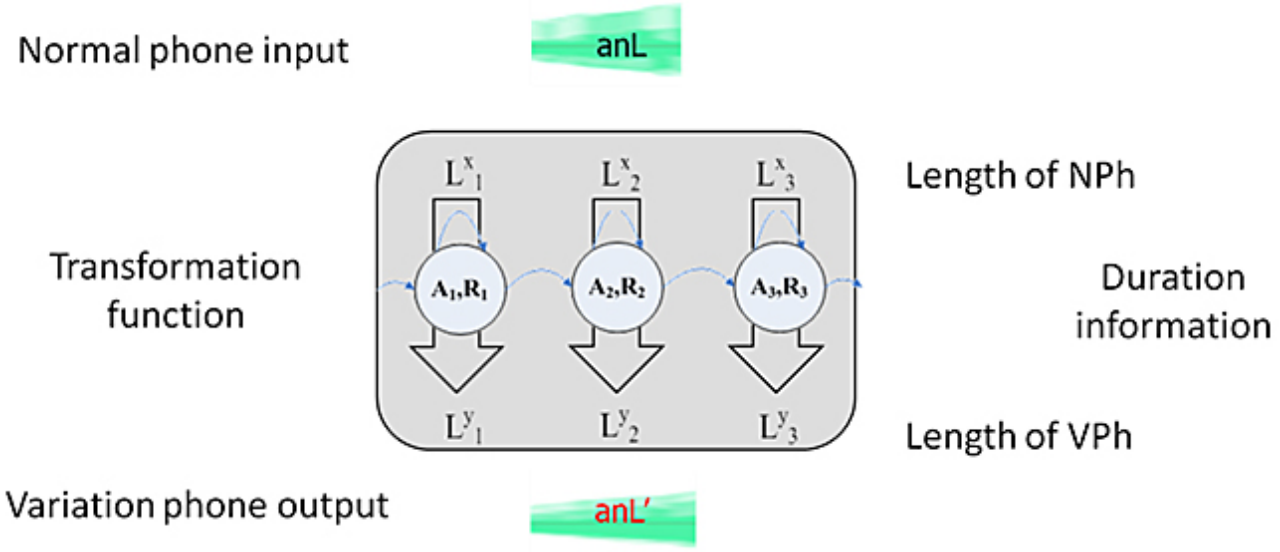

Figure 3 An illustration of the HMM-based duration model

Pronunciation Variation Prediction Model

The decision tree was adopted to model the relationship between the linguistic/articulatory features and the corresponding transformation functions. The transformation functions and duration scaling ratios are firstly clustered based on the Function-DT (F-DT) and Duration-DT (D-DT), respectively. The F-DT was then used to predict an appropriate transformation function for MGC coefficient and pitch conversion, and the D-DT was used to predict the state-based duration scaling ratio for duration conversion.

The training phase of the transformation function decision tree assigns each sample along with linguistic and articulatory features to a single leaf node. The mean of the generation errors in the F-DT is defined as:

$$GenErr_i = \sum_{m=1}^{M_i} \left\| \mathbf{y}_m - (\mathbf{A}_i \mathbf{x}_m + \mathbf{R}) \right\|^2 \quad (2)$$

where $\boldsymbol{y}_m$ is the m-th frame of the target speech *Y*, $\boldsymbol{x}_m$ is the m-th frame of the source speech X, $\mathbf{A}_i\mathbf{x}_m + \mathbf{R}$ is the linear transformation function of the *i*-th state, M is the total number of frames. In order to minimize the difference between the transformed feature vector of the source speech and the original feature vector of the target speech, the generation error reduction was defined, which is also known as variation.

As Fig. 4 shows, the spontaneous speech generated by the proposed method (Pro_S_VD) was preferred by the listeners in comparison with the speech synthesized using AVM, which synthesizes read speech of the target speaker. The results between Pro_S_VD and CMLLR show that listeners also preferred the proposed method. Some feedbacks given by the listeners show that the synthesized speech generated by the proposed method was slightly faster than that generated by the CMLLR method, and preferably regarded as spontaneous speech.

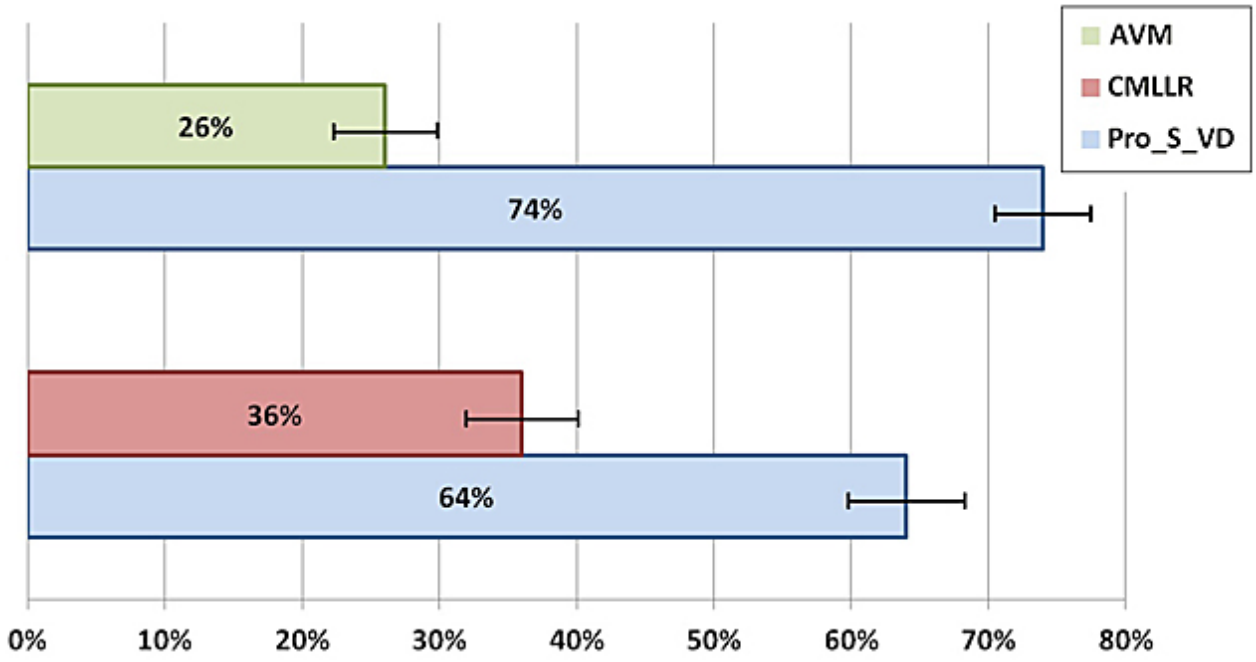

Figure 4 ABX preference test for speech spontaneity

This study presents an approach to improving the spontaneity in conventional HMM-based speech synthesis. Syllable contraction was modeled using a linear transformation function constructed from the parallel read speech and spontaneous speech with syllable contraction. Linguistic and articulatory features were utilized to categorize the transformation functions and construct a function decision tree (F-DT) and a duration decision tree (D-DT). The F-DT and D-DT select appropriate transformation functions for spectrum conversion and duration scaling using linguistic and articulatory features, respectively. Evaluation results indicate that the proposed transformation method can generate syllable contraction and improve apparent spontaneity with high naturalness and slightly inferior speech quality.

References

1. Tseng, S.-C., and Liu, Y.-F., “Annotation of Mandarin Conversational Dialogue Corpus,” *CKIP Technical Report*, No. 02-01, Academia Sinica, 2002.
2. Bennett C.L., and Black A.W., “Prediction of pronunciation variations for speech synthesis: A data-driven approach,” in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Philadelphia, Pennsylvania, 2005.
3. Werner S., Eichner M., Wolff M., and Hoffmann R., “Toward spontaneous speech synthesis - utlilizing language model information in TTS,” *IEEE Trans. Speech, Audio Processing*, pp. 436–445, 2004.
4. Sun L.-Y., and Wang Y.-R., “An Analysis Modeling of Syllable Contraction in Spontaneous Mandarin Speech Recognition,” Master Thesis, Dept. of Communication Engineering, NCTU, Taiwan, 2004.
5. Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S., “An Adaptive Algorithm for Mel-cepstral Analysis of Speech”, *in Proc. of ICASSP*, S7.11, PP. 453-456, 1991.