PHONE-BASED SPEECH SYNTHESIS WITH NEURAL NETWORK AND ARTICULATORY CONTROL

W. K. Lo and P. C. Ching

Department of Electronic Engineering
The Chinese University of Hong Kong
(Email : wklo@ee.cuhk.edu.hk & pcching@ee.cuhk.edu.hk)

ABSTRACT

This paper presents a novel method for synthesizing speech signal using a phone-based concatenation approach. Neural network is employed for the generalization of the phone templates during synthesis. Simplified articulatory space input parameters based on a modified vowel diagram are used to provide flexible and effective articulatory control. It also enables the design of an articulatory control model for allophonic variations in speech signal. The network approach is chosen for its non-linear mapping of the relationship between the articulatory space parameters and the spectral information of speech signal. In addition, non-linear approximation for phone template transitions is facilitated. The phone templates of the synthesizer are implicitly stored as network parameters of a medium size network. The performance of this new speech synthesis technique is demonstrated with a prototype system specifically designed for Cantonese (a common Chinese dialect) and the synthetic speech quality is assessed by informal listening tests.

1. INTRODUCTION

Speech synthesis is an area that tries to imitate the human speech production process. Research in this particular area is important in various aspects : i) for better understanding of human organism in the speech production mechanism; ii) for applications like speech prosthesis and human machine interfaces; and iii) for use as basis for the improvement in other research area like speech coding and recognition. Currently, there are several main stream techniques for speech synthesis : copy concatenation, parametric synthesis such as formant synthesis, and articulatory synthesis. Each of them has its own advantages and disadvantages.

From an application point of view, the intelligibility of most synthesis techniques are perceptually acceptable. However, satisfaction of this basic requirement does not lead to its wide spread application. The naturalness of the speech output is believed to be of great importance for its psychological effect on the users [1]. People may simply get unpleasant or tired and become reluctant to further usage. On the other hand, from the system implementation point of view, a more flexible but simple control is desirable and beneficial for better output. With the fast development of computing technology, the importance of storage requirement and system complexity is decreasing. However, for practical implementation, moderate storage and computation complexity are still desirable.

It is apparent that there are four main concerns for speech synthesis methodologies are : **controllability**, **naturalness**, **complexity** and **storage**. In general, controllability comes with higher complexity, and larger storage may give better naturalness. There is still not any single method that performs well on all these aspects. On the other hand, it is believed that as the technology continues to develop, articulatory synthesis will be the ultimate solution for the machine imitation of human speech production process.

2. NEURAL NETWORK SPEECH SYNTHESIS

The learning capability and non-linearity of artificial neural networks have found much application in speech synthesis. People have made use of these properties for text transcription and also for the mapping of vocal tract area function to digital filter parameters for synthesis purpose. The network approach has also found application to the inverse problem of mapping speech signal to vocal tract area function which is useful for articulatory synthesis. In this work, we will make use of the network approach to produce synthetic speech signal. Based on the presented inputs, a trained network will retrieve the appropriate phone templates and give generalized approximation for the transitions between templates. The whole synthesis process is under the control of articulatory type input parameters.

2.1. Synthesizer Network Architecture

The artificial neural network adopts a feed-forward architecture. It is trained by the log spectral magnitude of the phone templates. The selection of the phone templates tries to cover most of the vowel phonemes of the targeted language. At this moment, two different types of network has been tested.

Feed Forward Back-Propagation Network The common feed-forward back-propagation (FFBP) architecture is used. A three layer structure is adopted: an input layer, two hidden layers with sigmoidal activation function and an output layer with linear activation function. The sigmoidal function is used for incorporating non-linearity into the mapping and approximation. At the output layer, the linear activation function allows unbounded outputs for the log spectral magnitude of the signals. Figure 1 shows the FFBP structure of the synthesizer.

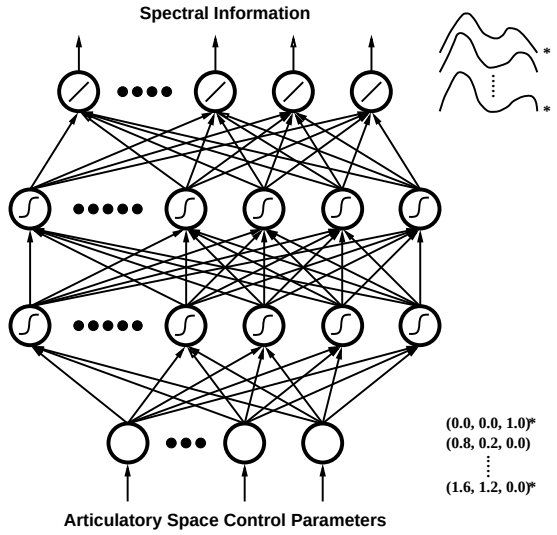

Figure 1: The FFBP network architecture for speech synthesis. It has three layers with two sigmoidal hidden layers and an linear output layer for spectral magnitude output.

Modified Itakura-Saito Distortion In addition to the sum squared error (SSE), a modified Itakura-Saito distortion (MISD) has also been used for training of the FFBP network. The MISD is defined as follows,

$$D = \frac{X}{\tilde{X}} - \frac{\tilde{X}}{X} + logX - log\tilde{X}$$

where X is the targeted spectra and $\tilde{X}$ is the synthesized spectra. The traditional Itakura-Saito distortion [2] has been widely used for speech processing. However, direct application of this function to neural network as error measure is not applicable (convergence problem). Therefore, modification of the distortion measure and using the MISD is necessary. As an illustration of the difference between MISD and SSE, figure 2 shows the training error convergence of the two distortion measure compared under MISD.

Radial Basis Function Network On the other hand, approximation network with radial basis function (RBF) has also been used as the synthesizer network. The RBF network has one input layer, one RBF layer and an linear output layer. Each RBF layer neuron output is given by

$$Z_{1,i} = exp\left(-C_{1,i} \cdot \sum_{j \in n2} W_{i,j} \cdot Z_{2,j}\right).$$

with $C_{1,i}$ being the RBF center, $n2$ the no. of input elements and $Z_{2,j}$ the output from input layer. The RBF network is trained with fixed spread using algorithm by Chen *et. al.* [3]. Basis function centers are chosen among the training templates. The potential of using this type of network and the training technique in the synthesizer network is that only the coverage of training templates is important but not their spread. Therefore, we may simply get a large amount of training templates and let the network selects itself. (Note : This formulation of approximation network is actually obtained from the solution of a regularization problem with Euclidean distance and differential regularization operator [4, 5].)

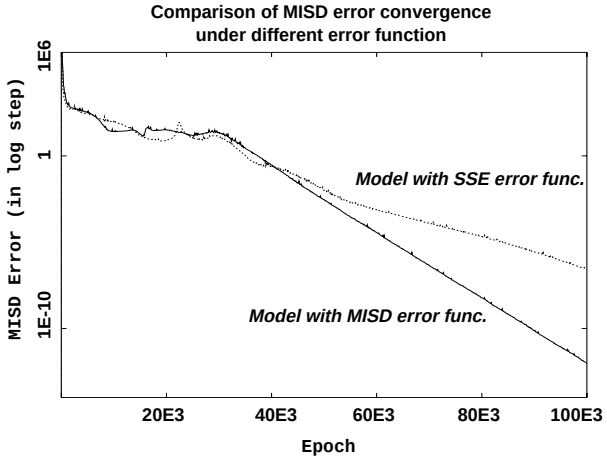

Figure 2: This figure shows that under the MISD error criterion, slower convergence is observed for the network trained with the SSE. The resulting synthesizer network also shows poorer speech quality.

2.2. Articulatory Control for the Synthesizer

Another characteristic of this synthesis methodology is that the concatenation synthesis is controlled by articulatory space parameters. This will give more control flexibility and also allow modeling of allophonic variation of speech signals through the variation of the articulatory control parameters. In addition, the variation in acoustic properties due to coarticulation can also be achieved by the introduction of controlled interference between adjacent sequence of control parameters for output utterance.

Vowel Quadrilateral In the current synthesis network, the simplified articulatory space control parameters are : **oral cavity openness**, **tongue body front-end position**

and **lip roundness**. These parameters are selected based on the cardinal vowel diagrams. The diagrams are modified and merged to form a three dimensional space that can represent most vowel sounds at different points within the space. The resultant articulatory space is illustrated in figure 3.

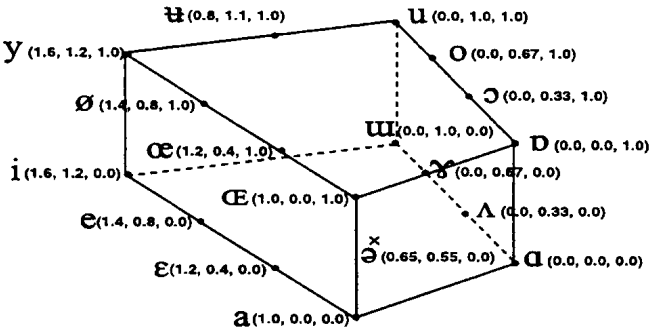

Figure 3: The modified vowel quadrilateral for the articulatory control space.

For the synthesis of speech utterances, a sequence of points within the articulatory space is first fed to the network. Each point essentially corresponds to a snapshot of the articulators position in time for producing the targeted speech. Upon presenting the sequence of control parameters, the corresponding sequence of spectra will then be returned by the network for the targeted speech.

In this method, the selection of the control parameters is not fixed. Although the three parameters selected are sufficient for most languages, enhancement can be made based on the target language, practical system requirement and training data availability.

3. A PROTOTYPE SYSTEM

A prototype system for Cantonese [1] speech synthesis has been constructed as a vehicle for the verification and evaluation of the proposed method [Figure 4].

Based on the sequential phonemic transcription input, the corresponding articulatory target points are retrieved from a database of target points. Together with appropriate timing information, the targets are interpolated to produce the appropriate control parameters for driving the synthesizer network. With the pitch and amplitude information, taking the inverse Fourier transform of the spectral information will produce the desired speech signal. In this prototype, synthesizer network of size ranging around 3-10-20-32 to 3-20-40-32 are empirically found to be better for a 8k Hz sampling rate. The training templates used are phone units cut from recorded utterance and the amount ranges from 11 to 20 of these cut out units.

[1] Readers may refer to [6, 7] for more detailed information on Cantonese.)

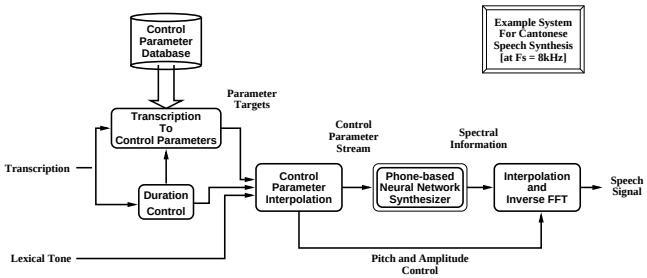

Figure 4: The system block diagram of a prototype synthesis system using the neural network synthesis method.

3.1. Articulatory Space Equivalence

In order to cater for practical implementation and simplification of the control parameter space, correspondences in the flat tongue vowel space have been assigned to nasal and plosive consonants. This agree fairly well with the actual manner of articulation in producing these sounds. Table 1 lists out their correspondence.

Target consonant	Flat tongue vowel correspondence
/n/	/i/
/m/	/u/
/ng/	/ɔ/
/t/	/i/
/p/	/u/
/k/	/ɔ/

Table 1: Correspondence of plosive and nasal consonants

One of the advantages of using the approach is that the control space of the synthesizer is much simplified. More importantly, this provides good transitions for the output signals between the consonants and their neighboring vowels. Whilst transition is known to be important for the quality of synthetic speech.

Two different cases will be considered in deriving the targeted consonants. Firstly, for nasal initials (onsets) and finals (rhymes), two synthesizer networks are used – one for nasal sounds and one for flat tongue vowel sounds. The sequence of control parameters are fed to each network during the corresponding segmental period. Their outputs are then overlapped and added together for a pre-defined transition period to obtain the final output signal. On the other hand, based on the correspondence list, plosive initials and stop codas etc are obtained with further control over the amplitude and timing of the signals. These correspondences have been tested and found to be able to successfully simplify the model.

4. RESULT

Table 2 shows the score of the synthesizer in an informal subjective listening test on a 0-10 scale (unacceptable – intelligible – fair – comfortable – human).

Vowel [a]	FFBP	RBF	FFBP-LPF [b]	RBF-LPF	LPC-10 [c]
/a6/	5.5	5	5.05	5.075	2.975
/i6/	3.75	3.5	2.75	3.625	2.425
/u6/	2.7	3.5	2.95	3.25	1.75
/ɔ6/	5	4.875	4.875	4.75	1.7
/ɛ6/	4.25	4.625	4.875	4.5	2
/œ6/	4.275	4	4	4.2	1.525
/y6/	3.5	3.25	4.375	4	1.375
/da3/	5.175	4.75	5.375	4.75	2.125
/di3/	4	4	2.75	3.5	2.625
/du3/	4	2.875	2.875	2.875	1.25
/dɔ3/	4	3.375	3.375	3.5	2
/dɛ3/	4.475	3.975	3.225	4.225	1.875
/dœ3/	3.225	2.975	2.725	3.725	1.25
/dy3/	3.975	4.125	3.125	3.5	2.125
/ai6/	7	6.875	7.875	7.625	4.5
/au6/	6.25	5	6.125	5.25	3.5
/ɔi6/	6.25	5.25	5.75	4.875	4
/ou6/	4.625	5.375	4.5	4.5	3.75
/ɸy6/	5.75	4.5	5.25	4.875	4
/ei6/	5.875	5.75	5.5	4	4.625
/ui6/	4.875	5	5.5	4.75	3
/iu6/	4.625	5.125	4.75	4.875	3.75
/ɔi6/	6.25	5.5	6.75	6.5	4.5
/ɔu6/	5.5	5.875	5.125	5.875	3.125

Table 2: The socre for subjective listening test on vowels, syllables and diphthongs.

[a] In the transcription, the numbers represent the tones of the syllables, upper-going as 3 and lower-going as 6.
[b] Low pass filtered version
[c] LPC-10 coded real speech

The speech samples are scored by five listeners (none expertise in speech processing). The main criterion is on assessing whether the speech samples sound synthetic or not. The result shows that the synthetic speech is of fair quality on average. As a reference, the synthetic speech samples all out perform the quality of LPC coded real speech. On the other hand, it can be observed that there is a trend that the quality increases from vowels, syllables to diphthongs. It is expected for speech listening test since short segments of speech are usually sound artificial for there is only few variations take place during the short period of time.

The diphthong of the FFBP and RBF used in the test are concatenated into two files and included as [SOUND A929S1.WAV] and [SOUND A929S2.WAV] respectively for reference (Fs=8kHz).

5. SUMMARY

Neural network approximation has been successfully applied to the problem of synthesizing speech signals. The articulatory control parameters give both simplicity and flexibility in the control process. Acoustic variations of speech signals due to allophonic variation or coarticulation can also be modeled by corresponding variations of the articulatory control parameters.

In summary, this method provides an intermediate degree of controllability when compared to articulatory synthesis and copy concatenation. On the other hand, the non-parametric templates give naturalness at training points from a segmental feature (acoustics) point of view. The small amount of phone templates are stored implicitly as the network parameters and take up little storage. Computation complexity of this method is mild for the simple network mapping and an IFFT operation. Moreover, the network approach has the potential of parallel implementation for distributing the computation load among provided resources.

6. ACKNOWLEDGMENTS

One of the authors, W. K. Lo, is gratified for the support of a studentship from the Croucher Foundation.

7. REFERENCES

1. Howard C. Nusbaum, Alexander L. Francis, and Anne S. Henly. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, 1:7–19, 1995.
2. F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan*, 53A:36–43, 1970.
3. S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2:302–309, 1991.
4. Tomaso Poggio and Federico Girosi. Network for approximation and learning. *Proceedings of IEEE*, 78:1481–1497, 1990.
5. S. Y. Kung. *Digital Neural Networks*. Prentice-Hall, 1993.
6. Oi kan Yeu Hashimot. *Studies in Yue dialects 1: Phonology of Cantones*. Cambridge University Press, 1972.
7. P. C. Ching, T. Lee, and Eric Zee. From phonology and acoustic properties to automatic recognition of cantonese. In *International Symposium on Speech, Image Processing and Neural Networks*, volume 1, pages 127–132, 1994.