SPEECH SYNTHESIS USING ARTIFICIAL NEURAL NETWORKS

E. Veera Raghavendra†, *P. Vijayaditya*†, *Kishore Prahallad*†‡

†International Institute of Information Technology, Hyderabad, India.
‡Language Technologies Institute, Carnegie Mellon University, USA.

raghavendra@iiit.ac.in p.vijayaditya@gmail.com skishore@iiit.ac.in

ABSTRACT

Statistical parametric synthesis becoming more popular in recent years due to its adaptability and size of the synthesis. Mel cepstral coefficients, fundamental frequency (f_0) and duration are the main components for synthesizing speech in statistical parametric synthesis. The current study mainly concentrates on mel cesptral coefficients. Durations and f_0 are taken from the original data. In this paper, we are attempting on two fold problem. First problem is how to predict mel cepstral coefficient from text using artificial neural networks. The second problem is predicting formants from the text.

Index Terms: speech synthesis, formants, statistical parametric speech synthesis.

1. INTRODUCTION

Parametric speech synthesizers in early 80's, also referred to as synthesis-by-rule, were built using careful selection of parameters and a set of rules for manipulation of parameters. Statistical Parametric Synthesis (SPS) uses machine learning algorithms to learn the parameters from the features extracted from the speech signal [1]. HTS [2, 3] and CLUSTERGEN [4] are statistical parametric synthesis engines using hidden Markov models and Classification and Regression Trees (CART) respectively to learn the parameters from the speech data. In SPS framework, spectral features are often represented by Mel-Log spectral approximation based cepstral coefficients, line spectral pairs and harmonic noise models features. Excitation features are represented by fundamental frequency and voicing strengths. Source-filter models are used to generate speech signal from excitation and spectral features [5].

In this work, we propose two methodologies for synthesizing speech using artificial neural networks. The first method is predicting Mel-Cepstral Coefficients and synthesize speech using MLSA [5] vocoder. The second method is building a statistitical parametric synthesis using formant features. The need for such an investigation lies in the fact that formants are more flexible parameters than cepstral coefficients. Formants allow simple transformation to simulate several aspects of voice quality, speaker transformation etc., and also on the other hand our understanding of speech production mechanism is better in terms of formants and their bandwidths [6]. While many of the early rule based synthesizers used formants to synthesize speech, the current investigation differs from these earlier works as the formants and bandwidths extracted from the speech signal are used to train parameters of machine learning models which are capable of predicting the formants from the text directly during synthesis phase. Moreover, the rules required to incorporate coarticulation, and natural variations of formants within a phone are also being learnt automatically.

2. DATABASE USED

In all the experiments reported in this paper, RMS voice from CMU ARCTIC dataset was used. Out of 1132 utterances, 1019 utterances were used as a part of training and the remaining utterances were used for testing. The parameters extracted from the speech signal were 25 coefficient Mel-Cepstral Coefficients (MCEPs), seven formants, seven bandwidths, and f_0 using ESPS formant extraction [7] and one energy value with 25 milliseconds frame size and 5 milliseconds frame shift.

3. OVERVIEW OF THE ANN BASED SYNTHESIS

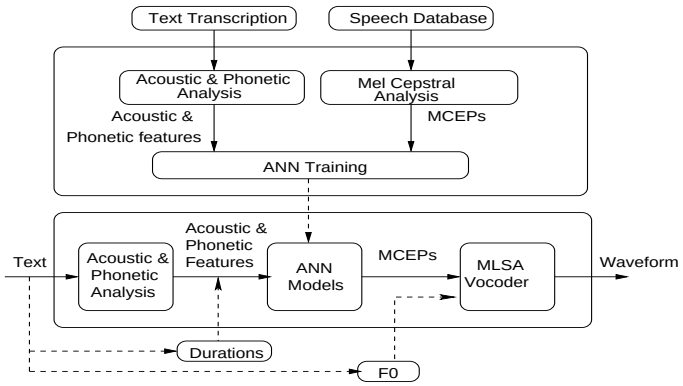


Fig. 1. Mel Cepstral based ANN synthesis architecture

The complete system is shown in Fig 1. The text-to-speech system includes a text-to-acoustic&phonetic analysis subsystem, one/more neural network models used to predict MCEPs. During synthesis, the given text is converted into acoustic & phonetic notation and MCEPs are predicted using existing models. Here the durations of each phoneme for and fundamental frequencies (f_0) are taken from the test sentence database. Predicted MCEPs and original f_0 are given to MLSA vocoder to synthesize speech.

4. ARTIFICIAL NEURAL NETWORK MODELS FOR SPEECH SYNTHESIS

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. For example, a feedforward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. ANN models are also known to capture complex and nonlinear mapping and for their generalization behavior. In the context of speech synthesis, a mapping is required from text (linguistic

space) to speech (acoustic space). Thus we exploit the pattern mapping capabilities of ANN models to perform complex and nonlinear mapping of linguistic space to acoustic space to generate synthetic speech.

4.1. Input Representation of Text

Since we are performing a mapping from text input space to formant output space, a careful representation is needed at the input layer as such mapping is not only complex but we also expect the ANN model to produce subtle variations in the formants and bandwidths for every frame.

Features extracted from the text to train the ANN model is shown in Table 1. The features include current, left and right phone articulatory and syllable features. Along with these, current phone position in the word, current word position in the sentence and temporal features (position of the frame within the current phone) and state information of the current frame. Please note that the state information is incorporated in the ANN modeling to help to differentiate the frames within a phone. To represent the temporal variations, fifteen time index neurons are used within a state, following formula [8] is used. These time indices represent the relative position of the current frame within a state of a phone segment. This helps to smooth transition between neighboring frames especially on state and segment boundaries. The value of time index i during frame j is calculated using Eq 1 (we have chosen β = 0.01), such that time index i reaches its maximum value during frame $j = i$.

$$O_i = exp(-\beta(i-j)^2) \quad (1)$$

4.2. Output Representation

The network is expected to predict Mel Cepstral Coefficients (MCEPs) at the output layer. 25 coefficient vector is predicted for each 10ms frame size with 5ms frame shift. The ANN model is trained for 200 iterations using back propagation learning algorithm.

5. EXPERIMENTS WITH ANN MODELS

5.1. One network for all the phones

The purpose of this neural network is to generate MCEPs. Features mentioned in 1(a) are used to represent mapping between input and output. Generally statistical models require huge amount of data to analyze the training patterns. Hence, we wanted to build model with only one network. The architecture of the feedforward network used in this work is a five layer network: 136 L 75 N 25 S 75 N 25 L, where the numbers indicate the number of nodes in the corresponding layer. L represents linear activation function, N represents tangential activation function and S represents sigmoid activation function. Fig. 2 shows the error curve of the ANN model obtained during training. The monotonically decreasing error curve demonstrates that it is possible to train an ANN model to perform complex mapping from text to formants.

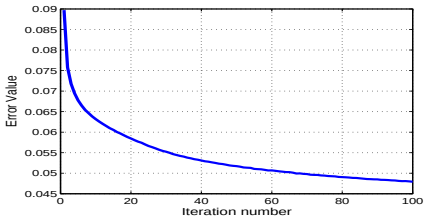


Fig. 2. ANN training error curve for one network for all the phones

Table 1. *Input features to predict formants and bandwidths*

(a) Overall features to map between input and output

Feat. Name	# Bits	Rep.
Current phone articulatory features	29	Binary
Previous phone articulatory features	29	Binary
Next phoneme articulatory features	29	Binary
Current phone position in the word	3	Binary
Current phone syllable features	9	Binary
Previous phone syllable features	9	Binary
Next phone syllable features	9	Binary
Current word position in sentence	3	Binary
Temporal features	15	Float
Phone Duration	1	Float
Total	136	

(b) Articulatory features used in the input

Feat. Name	Feat. Values	# Bits
Phone type	Vowel/Consonant	2
Vowel length	Short/long/diphthong/schwa	4
Vowel height	High/middle /low	3
Vowel frontness	Front/mid/back	3
Lip rounding	+/-	2
Consonant type	Stop/fricative/affricative/ nasal/lateral/approximate	6
Place of articulation	Labial/alveolar/palatal/ labio-dental/dental/velar/glottal	7
Consonant voicing	voiced/unvoiced	2

(c) Syllable features used in the input

Syllable Feat.	Feat. Values	# Bits
Stress	True/false	1
Phone type	Onset/coda	2
Phone position in syllable	Begin/middle/end	3
Syllable position in word	Begin/middle/end	3

(d) Other features used in the input

Feature name	Feat. Values	# Bits
Phone position in the word	Begin/middle/end	3
Word position in the sentence	Begin/middle/end	3
Phone state information	Begin/middle/end	3

To evaluate synthesis quality, Mel Cepstral Distortion (MCD)[9] is computed on held-out data set. The measure is defined as

$$MCD = (10/ln10) * \sqrt{2 * \sum_{i=1}^{25} (mc_i^t - mc_i^e)^2} \quad (2)$$

where mc_i^t and mcd_i^e denote the target and the estimated mel-cepstra, respectively. MCD is calculated over all the MCEP coefficients, including the zeroth coefficient. Lesser the MCD value the better it is, and informally we have observed that a difference of 0.2 in MCD value produces difference in the perceptual difference in quality of synthetic speech. The MCD value we obtained is 6.47.

5.2. Separate network for vowels and consonants

We informally observed that there is some problem in mapping when all types of phones are combined. Hence, the data is separated into two parts; vowels and consonants, and one network is built for each type. Though training data is less comparatively with previous experiment, the mapping would be easy. The architecture of the feedforward network used in this work is a five layer network: 136 L 75 N 25 S 75 N 25 L.

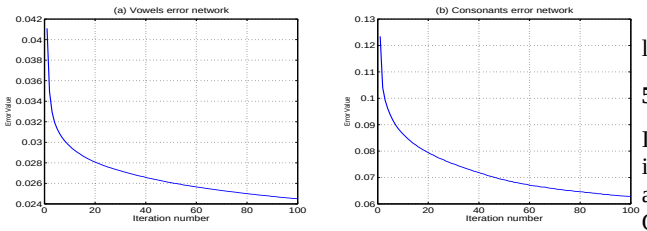


Fig. 3. ANN training error curves for vowels and consonants

Fig. 3 shows the error curve of the ANN model obtained during training. We can observe that, vowel type network error has decreased tremendously comparing to the previous experiment but consonant type network is little higher than previous experiment. The objective measure MCD also decreased 0.002 and the value is found to be 6.45. It show that multiple networks gives better result.

5.3. Separate network for each state

Based on above experiments we further divided the data into further level. Instead of using phones as the smallest unit, we considered $state$ as the basic unit. State level segments are obtained from the EHMM[10] segmentation. EHMM considers three states for each phone; staring, middle and ending. The architecture of the feedforward network used in this experiment is a five layer network: 136 L 75 N 25 S 75 N 25 L. In this experiment MCD value is drastically change from 6.45 to 5.87. We observed that state based modeling would be better choice.

5.4. One network for all the states

In this experiment we wanted to experiment with one network for all the states. But, some how we want to represent state information in the input features otherwise there will not be any difference with first experiment. Hence, we have represented state information with three bits for each frame as shown in 1(c) and state level duration is used as duration value. To vary among the state segments we also introduced f_0 as one more feature. The architecture of the feedforward network used in this experiment is a five layer network: 140 L 75 N 25 S 75 N 25 L. This has given similar MCD value as we got in previous experiment. The MCD value is found to be 5.86. From this experiment we observed that more contextual information at state level would be more useful for network mapping. It means that more variations from frame to frame is better.

5.5. Experiments with different architectures

We know that network architecture also plays a vital role in the performance of the synthesis. Hence, we have experimented with multiple architectures. Table 2 show the MCD values for each architecture.

Table 2. *ANN network architectures and corresponding MCD values.*

Architecture	MCD
140 L 100 N 25 S 100 N 25 L	5.87
140 L 100 N 15 S 100 N 25 L	5.85
140 L100 N 10 S 100 N 25 L	5.9
140 L 100 N 6 S 100 N 25 L	5.94
140 L 210 N 15 S 210 N 25 L	5.81

From above table we can observe that 1.5 time nodes of the input layer in second and fourth layer gives the better results.

5.6. Applying MLPG on predicted MCEPs

Informal studies showed that speech produced by above technique is understandable but not natural. The voice appears as robotic. To alleviate this problem we have used Maximum Likelihood Parameter Generation (MLPG) [11] to obtain smoother trajectories. The MCD value obtained to be 5.74. Section Section 5.7 gives the difference between ANN synthesis with MLPG and with out MLPG.

5.7. Experiment

So far we discussed all the experiments with the help of objective evaluation. To evaluate the synthesizers perceptually we have conducted subjective evaluation between CLUSTERGEN [4] and ANN synthesizers; with out MLPG and with MLPG, discussed in above sub sections. The subjects participated in this study are non-native speakers of English but all of them are graduated students. For these experiments we have selected 10 utterances from test database. The subjects participated in these tests do not have any experience in speech synthesis. Each listener is subjected to Mean Opinion Score (MOS) i.e score between 1 (worst) to 5 (best) and AB-Test i.e the same sentence synthesized by two different synthesizers is played in random order and the listener is asked to decide which one sounded better. They also had the choice of giving the decision of equality.

6. FORMANT PREDICTION USING ANNS

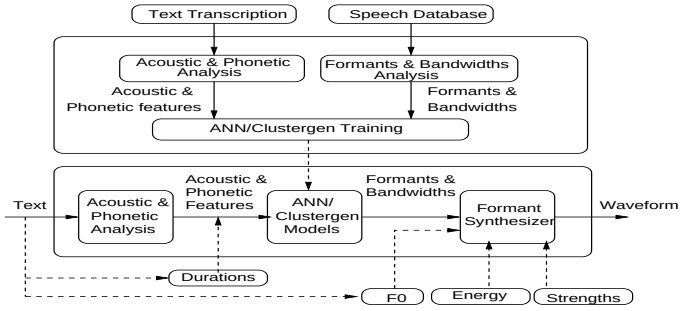


Fig. 4. Formant based synthesis architecture

In the previous section we discussed prediction of MCEPs from the text using ANNs. In this section, we describe a method of building a statistical parametric synthesizer using formants as parameters. The need for such an investigation lies in the fact that formants are more flexible parameters than cepstral coefficients. Formants allow simple transformation to simulate several aspects of voice quality, speaker transformation etc., and also on the other hand our understanding of speech production mechanism is better in terms of formants and their bandwidths. Figure 4 show the architecture of formant based synthesis architecture with ANN and CLUSTERGEN. Next section discuss more about the CLUSTERGEN based formant prediction. Klatt [12] used 39 features for synthesizing speech. The author extracted all the features manually very carefully from the speech signal. As it is very difficult to extract all those features with current technology, we are experimenting only with formants as a first step. Energy, strengths and f_0 are used from original data. The formants are predicted using ANN with same input features mentioned in 1. As output we have used 14 coefficient; 7 formants and 7 bandwidths, vector. Since there is wide difference between each formant of the frame, all the columns are normalized with mean and variance of each column.

7. FORMANT PREDICTION USING CLUSTERGEN

CLUSTERGEN is a SPS engine using CART models to predict the acoustic features from the given input text. While the framework of CLUSTERGEN is flexible, it typically uses Mel-cepstral coefficients derived from Mel-Log Spectral Approximation (MLSA) technique. In this work, the CLUSTERGEN was adapted to predict formants and bandwidths from the text. The standard build process of CLUSTERGEN was used to build the RMS voice using formants and bandwidths. CART trees are built by finding questions that split the data to minimize impurity in the cluster. At each leaf node, a mean vector is derived as a representation of the cluster of units.

8. SYNTHESIS FROM FORMANTS

To synthesize speech from formants we adapted two different strategies. The first method is conventional form of synthesis where formants are converted into linear prediction coefficients and speech is synthesized using source-filter model. The second method is to perform another transformation from formant space to cepstral coefficient space and the speech is synthesized using MLSA synthesis technique. ESPS [7] toolkit was used for formant extractions.

8.1. Method I

The formant frequencies F_k and their bandwidths B_k, where k denotes the formant index, can be used to derive the roots of the prediction polynomial/poles using the equation 3.

$$\theta_k = \frac{2\pi F_k}{f_s} \quad and \quad \rho_k = e^{(\frac{-B_k 2\pi}{f_s})} \tag{3}$$

where ρ_k and θ_k are the pole radius and the normalized formant frequency respectively. These roots are used to derive the linear prediction polynomial coefficients [13].

LPC synthesis equation is used to generate speech from the prediction polynomial. The control parameters in formant synthesis are normally updated every 5ms for mimicking the rapid formant transitions and brief plosive bursts [12]. However LPC parameters held too long during the production of speech give the perception of buzzy quality. To avoid this, the lpc parameters are interpolated for every sample. In order to maintain the stability of the LPC synthesizer the predictor coefficients are converted into log area coefficients prior to interpolation [14].

To further reduce the buzzy quality of speech, mixed excitation is used. This excitation method uses different mixtures of pulse and noise in 5 frequency bands, where the relative pulse and noise mixtures are derived from the band pass voicing strengths of the five frequency bands for every frame [15]. The LF model for differentiated glottal pulse was used to model the glottal source signal and the lip radiation [13]. The source was generated using mixed excitation model [15]. As stated in the paper the strengths and f0 are required for generating the residual. The radiation characteristic adds a gradual rise in the overall spectrum [12]. However the parameters of the LF model were kept constant across the duration of the sentence and also across the speakers. When formants and bandwidths are used for synthesis, ringing noise is perceived if the appropriate glottal roll-off is not provided. To alleviate this mixed excitation output was passed through a filter modelling the glottal source. LF model for differentiated glottal pulse was used to model the glottal source signal and the lip radiation [13].

8.2. Method II

The second method is to perform another transformation from formant space to cepstral coefficient space. Such transformation is done through the use of another artificial neural network (referred to ANN-2). The input to this ANN-2 are the formants and bandwidths as predicted in Section 3 or 4, and the output are the Mel-cepstral coefficients corresponding to that frame. This network could be viewed as nonlinear transformation of formants to cepstral coefficients, and also could be viewed as error correction network. The effect of any error in the prediction of formants could be minimized in the transformation process. The generated Mel-cepstral coefficients are used to synthesize speech using MLSA synthesis technique. In order to conduct a objective analysis, Mel-cepstral distortion was computed for samples from Method II, and it was found to be 6.14.

9. EVALUATION OF PREDICTION OF FORMANTS

9.1. Visual Representation of Formant trajectories

Figure 5 shows the first, second, and third formant frequencies for the word *gregson*. A comparison is made between formants extracted from the original speech signal, formants predicted from ANN models and formants predicted from CLUSTERGEN. While both ANN models and CLUSTERGEN in general are able to produce required trajectories, the ANN models seem to produce smoother trajectories than CLUSTERGEN which could be attributed to the generalization abilities of ANN.

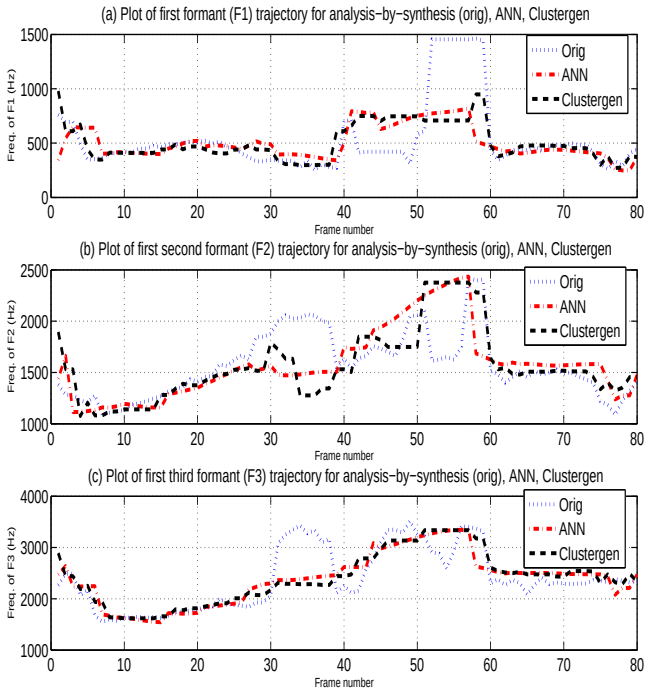
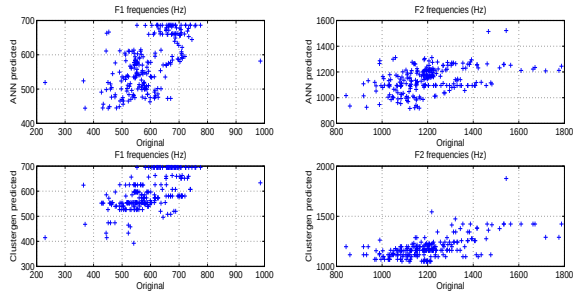


Fig. 5. Visual representation of F1, F2, and F3 formants in original, predicted from ANN and CLUSTERGEN.

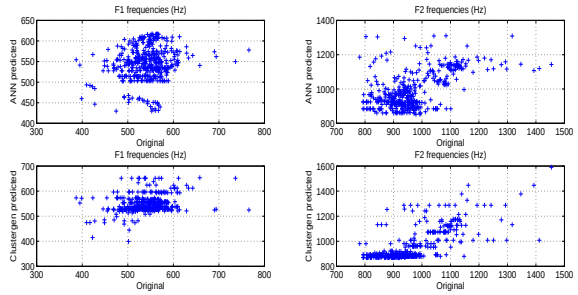
9.2. Analysis of formants in vowels

Fig. 6 shows the scatter plot displaying the correlation between the formants extracted from the speech signal and the formants predicted

from CLUSTERGEN and ANN. It could be observed there exists sufficient correlation between the original formants and the predicted ones. The formants predicted from CLUSTERGEN may look fewer in number for the reason that CLUSTERGEN is predicting same frequencies over number of frames in a particular state.



(a) Phoneme /aa/



(b) Phoneme /ao/

Fig. 6. Measured frequencies of first and seconds formants for vowels.

9.3. Root mean squared error

In order to evaluate, objectively, the prediction accuracy between predicted and original values of test sentences, root mean squared error (RMSE) is calculated. RMSE is calculated separately for each formant.

$$d = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2 / N} \tag{4}$$

Where d is the root mean square error, x_i is the original value, y_i is the predicted value and N is the number of examples. Objective evaluation of the deviations of each formant are given in the Table 3. From Table 3 we can observe that ANN and CLUSTERGEN are able to predict efficiently in vowels and first formants and other needs more prediction accuracy in other formant regions and consonants.

10. EVALUATION OF SYNTHESIZED SPEECH

The speech synthesized from Method I and II, was perceived to be intelligible in some informal experiments. However, the quality of signal from Method II was found to be smoother than Method I. It was observed that the excitation signal in Method I needs to be improvised for a smoother and better sounding quality speech. Speech samples for Method I and II are available in the following link: http://ravi.iiit.ac.in/śpeech/samples/icon-09/.

Table 3. *Root mean squared error between ANN predicted and analysis-by-synthesis formants and CLUSTERGEN predicted and analysis-by-synthesis formants. F1 denotes first formant, F2 denote second formant, F3 denotes third formant, F4 denote fourth formant, F5 denotes fifth formant, F6 denotes sixth formant,V denotes Vowel and C denotes Consonants.*

	ANN		CLUSTERGEN	
	V	C	V	C
F1	90	259	75	238
F2	206	319	147	297
F3	224	332	208	321
F4	266	365	259	363
F5	323	371	323	373
F6	283	305	287	307
F7	319	269	321	268

11. REFERENCES

[1] A. Black, H. Zen, and K Tokuda, "Statistical parametric synthesis," in *Proceedings of ICASSP*, 2007, pp. IV–1229–IV–1232.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and Kitamura T., "Speech parameter generation algorithms for hmm-base speech synthesis," *in Proceedings of ICASSP*, 2000.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," *in Proceedings of ICASSP*, pp. vol. 1, pp. 660–663., May 1995.

[4] Black, A., "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling," in *Proceedings of Interspeech*, 2006, pp. 1762–1765.

[5] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of ICASSP*, 1983, pp. 93–96.

[6] D.O'. Shaughnessy, *Speech Communication*, 2004.

[7] ESPS, "Esps source code from the esps/waves+ package," 2009, [Online; accessed 9-April-2009].

[8] C. Fatima and G. Mhania, "Towards a high quality arabic speech synthesis system based on neural networks and residual excited vocal tract model," *Signal, Image and Video Processing*, vol. 2, no. 1, pp. 73–87, January 2008.

[9] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 31–36.

[10] K. Prahallad, A.W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proceedings of ICASSP*, France, 2006.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM based speech synthesis," *International Conference on Acoustics, Speech and Signal Processing*, June 2000.

[12] D.H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, March 1980.

[13] P. Satyanarayana, *Short segment analysis of speech for enhancement*, Ph.D. thesis, IIT-Madras, 1999.

[14] M. Sambur, A. Rosenberg, L. Rabiner, and C. McGonegal, "On reducing the buzz in lpc synthesis," in *Proceedings of IEEE*, May 1977, vol. 2, pp. 401–404.

[15] A.V. McCree and III Barnwell, T.P., "A mixed excitation lpc vocoder model for low bit rate speech coding," *IEEE Transaction on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.