AN EXPRESSIVE HMM-BASED TEXT-TO-SPEECH SYNTHESIS SYSTEM UTILIZING GLOTTAL INVERSE FILTERING FOR TAMIL LANGUAGE

Sudhakar B. and Bensraj R.
Department of Electrical Engineering, Annamalai University, Tamil Nadu, India
E-Mail: balrajsudhakar@gmail.com

ABSTRACT

This paper describes an Hidden Markov Model (HMM) based speech synthesis system that make use of the Glottal Inverse Filtering (GIF) for producing natural sounding synthetic speech in Tamil language. GIF based method is used for parameterization. Tamil speech is first parameterized into spectral and excitation features in the proposed system. The HMM system is trained by utilizing the speech parameters and then generated from the trained HMM according to the given Tamil text input. In this proposed work the voice sources are glottal flow pulses extracted from real speech, and the voice source is additionally customized according to the all-pole model parameters produced by the HMM. Experimental results show that the proposed system is accomplished of generating natural sounding speech, and the quality is obviously better compared to a system exploiting a conventional impulse train excitation model.

Keywords: Tamil TTS, GIF, HMM.

1. INTRODUCTION

The ultimate goal of text-to-speech synthesis (TTS) is to facilitate generating natural sounding speech from impulsive text. Moreover the existing trend in TTS investigation is capable of constructing speech in diverse speaking styles with different speaker characteristics and also analyses sentiments of different speakers. In order to accomplish these rigorous universal requirements, the most important synthesis techniques have concerned escalating interest in the speech research community. The unit selection technique and HMM based approaches are the two most important techniques to improve the performance of TTS synthesis for nature sounding. The unit selection techniques do not permit for adjustment of the TTS system to diverse speaking styles and speaker characteristics and requires databases of broad sizes. The HMM based techniques are promoted from better adaptability and evidently lesser memory requirement. The TTS synthesis system is to search for a unique method aiming at accurate modeling of different voice characteristics as well as prosodic features of speech. The HMM based synthesizers have been developed with special emphasis on voice characteristics such as speaker individualities, speaking styles, and emotions [1]. Statistical parametric speech synthesis based on HMM [2] are demonstrated very effectively in synthesizing acceptable speech. A novel approach has been demonstrated on speech synthesis approach in [3]. The utility and effectiveness of linear regression algorithms (LRA) in speaker adaptation for HMM based speech synthesis is presented in [4]. A novel approach to voice characteristic conversion for HMM based text-to-speech synthesis system by using speaker interpolation is introduced in [5]. Quantized Hidden Markov Models (QHMM) is applied to automatic speech recognition in embedded devices without loss of recognition performance is described in [6]. A speaker adaptive HMM based speech synthesis system is presented in [7]. The possibility of using complex cepstrum for glottal flow estimation on a large scale data base is investigated [8] by the zeros of the Z-Transform (ZZT) technique. A comparative study has been carried out [9] in the basic problems of speech processing. An improved glottal source built on the Liljencrants-Fant model (LF) and its effectiveness in speech synthesis systems are presented in [10]. The use of LF model to represent the glottal source signal to HMM based speech synthesis system is prescribed in [11]. A novel synthesis speech stimuli is introduced [12] by comparing artificial and biological Neural Networks. An HMM based speech synthesis for GIF for generating natural sounding synthetic speech is proposed in [13].

In this paper, a HMM-based Tamil speech synthesis system that exploits GIF for producing natural signaling synthetic speech for Tamil language is presented. The paper is organized as follows: section 2 describes the HMM based speech synthesis system. The proposed HMM based TTS system is presented in section 3. The experimental results and discussion on the proposed speech synthesis system and future plans are presented in section 4 and conclusion is presented in Section 5.

2. HMM BASED SPEECH SYNTHESIS SYSTEM

The HMMs principal speech synthesis develops the identical modeling logic as in speech recognition namely, representing speech as a constrained sequence of random observations characterized by their second-order statistics. Substantial differences from the HMMs used in speech recognition include the following.

- The clear explanation of the pitch, by adding the log-scaled fundamental frequency (log-F0) and its first and second order derivatives to the usual feature vector of Mel-Frequency cepstrum Coefficients (MFCCs) which describes the spectral envelope. The use of Multi-Space Density functions, to accommodate a discrete voiced/unvoiced decision variable observed in conjunction with the continuous log-F0 values.

- The definition of so-called full-context models, which expand the n-phones with a richer set of context descriptors that go beyond the co-articulation effects, and which are related to the lexical or syntactic levels of the training sentences. This entails a combinatorial increase of the number of context dependent models, and a problem of scarcity of the training data available for each model. This problem is tackled by the application of standard tree-based state clustering techniques.
- A separate state duration model is trained for each context dependent model on the basis of the state occupancies over the training set.

In speech recognition, the trained models are used as templates to be matched via the likelihoods of incoming observations. Conversely, for synthesis, a speech parameter generation algorithm is applied to emit some smooth sequences of synthetic MFCC and log-F0 features, in Maximum Likelihood accordance with a selected sequence of states.

3. PROPOSED SPEECH SYNTHESIS SYSTEM

The proposed HMM Tamil TTS system targets to produce natural sounding synthetic speech proficient of carrying diverse styles of speaking and also emotions [14]. To accomplish this objective, the task of the real human voice construction machine is modeled by utilizing GIF entrenched in an HMM framework. We can compute the source of voiced speech and glottal volume velocity wave form from the speech pressure signal by using GIF. The below equation is used to estimate the glottal volume velocity G(z) of GIF .

$$G(z) = \frac{Ss(z)}{Vt(z)\, Lr(z)} \qquad (1)$$

Where Ss(z), Vt(z) and Lr(z) represents the Z-Transform of Speech signal, Z-Transform of Vocal tract and Z-Transform of Lip radiation effect. Parametric feature expression for the voice source and the vocal tract transfer function are computed in the parameterization phase using automatic GIF. The Vocal tract transfer function (VTTF) normally consist of poles and zeros and can we expressed as

$$V(z) = \frac{b_0 \prod_{k=1}^{m} (1 - d_k z^{-1})}{\prod_{k=1}^{n} (1 - c_k z^{-1})} \qquad (2)$$

Where b_0, c_k and d_k represents gain factor, poles of V(z) and zeros of V(z). The poles represents several peaks associated to resonance of the acoustic cavities that from the vocal tract. These resonance or measured by formants. Each formant is described by its formant frequency and its formant bandwidth. The zeros or anti-resonance of the VTTF represent energy loss and located at very high frequencies. During the synthesis phase, natural glottal pulses are utilized for producing the source signal for voiced sounds. To reproduce the time-varying variations in the real voice source, the spectral envelope of this glottal excitation waveform is changed with an adaptive Infinite Impulse Response (IIR) filter. The present application of the system is applied for the Dravidian language Tamil.

The proposed system comprises of two main parts: training and synthesis. In the training phase, speech parameters calculated by GIF are take out from sentences of a Tamil speech database. The acquired speech parameters are modeled in the HMM framework. In the synthesis phase, the HMMs are concatenated according to the analyzed input text and speech parameters are generated from the HMM. Then the parameters are given into the synthesis module for generating the speech waveform.

4. PARAMETERIZATION PHASE

In order to eliminate possible low-frequency fluctuation from the signal, the signal is high pass filtered first in the parameterization stage. GIF requires integration since high-pass filtering is important. The signal is then windowed with a rectangular window to 25-ms frames at 5ms [13]. The speech features from each frame is shown in the Table-1. It depends on the speaker and typically lower order spectral models work sound for female while greater order produce superior results for male.

Table-1. Speech features and its parameters.

Features	Number of parameters
Fundamental frequency	1
Spectral energy	5
Voiced spectrum	20
Unvoiced spectrum	20

The log-energy is estimated from the windowed speech signal. Then GIF is accomplished in order to evaluate the glottal volume velocity waveform from the speech signal. Iterative Adaptive Inverse Filtering (IAIF) is employed for the automatic GIF. It repeatedly withdraws the effects of the vocal tract and the lip radiation from the speech signal using all-pole modeling. The assessed glottal flow signal and the Linear Predictive Coding (LPC) model of the vocal tract are the outputs of the inverse filtering block. In order to arrest the deviations in the glottal flow due to different phonation or speaking style, the spectral envelope of the glottal flow is further parameterized with LPC.

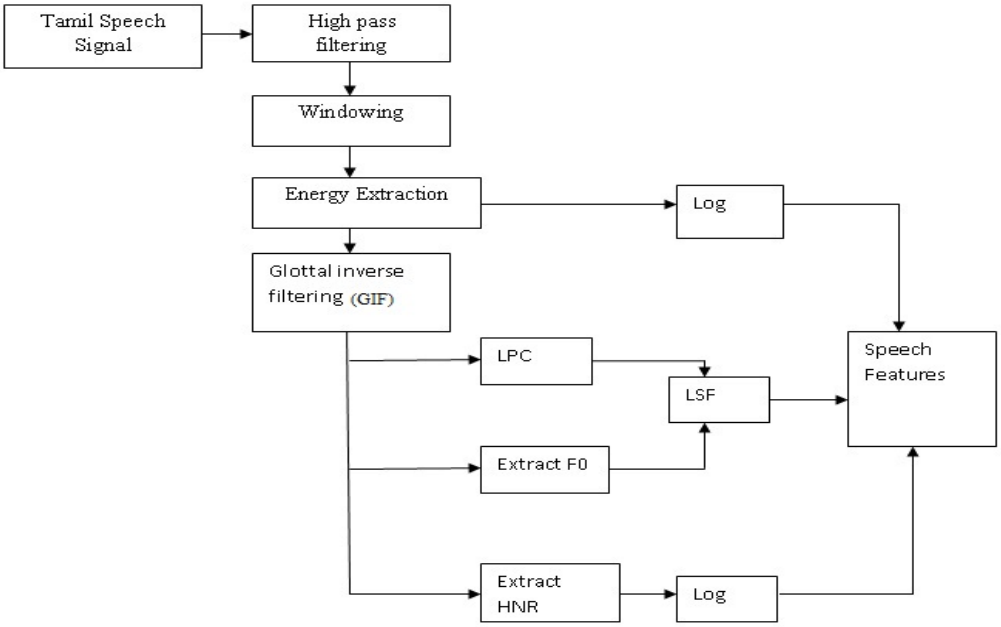

Figure-1. Block diagram of Parameterization phase.

As per Figure-1 the Tamil speech signal is decomposed in to the glottal source signal G(n) and all pole model of vocal tract V(z) utilizing the IAIF method. The G(n) is again parameterized into the all pole model of voice source G(z), the fundamental frequency (F0) and harmonic-to-noise ratio (HNR). The result and parameters or converted to appropriate representation for HMM system. As well, a LPC model is calculated for silent speech sounds straight from the speech frame. All the LPC models are transformed to Line Spectral Frequencies (LSF). It is a parametric demonstration of LPC information compatible to be used in a statistical HMM system. Voiced and unvoiced spectrums of LSF are additionally converted to the Mel-Frequency scale.

Using the autocorrelation method, fundamental frequency is estimated from the glottal flow signal in order to estimate the degree of voicing in the glottal flow signal. A harmonic -to- noise ratio (HNR) is defined with respect to the ratio between the upper and lower smoothed spectral envelopes. It is the average of the five frequency bands are rendering to the corresponding rectangular bandwidth scale. Conventional LPC is used to evaluate the spectral model of speech in case of unvoiced speech.

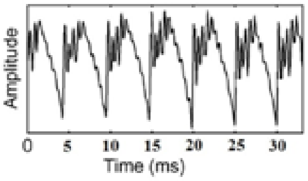

Figure-2. Waveform of a Tamil vowel [m] without GIF.

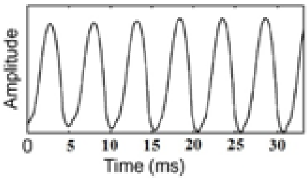

Figure-3. Waveform of a Tamil vowel [m] using GIF estimated by IAIF method.

The Waveform in Figure-2 generated by a male speaker without using GIF has been illustrated. In this we are facing more noises and distortions. So we are not able to accurately calculate the speech parameters. In Figure-3 Waveform has been generated by a male speaker using GIF estimated with the (IAIF) method. In this condition the glottal flow pulses are generated without distortion. So we can easy to calculate the speech parameters from Figure-3 when we compare with Figure-2.

The excitation signal involves of voiced and unvoiced sound bases. Glottal flow pulse pulled out from a natural vowel is the basis of the voiced sound source. The usage of an actual glottal pulse helps in stabilizing the naturalness and quality of the synthetic speech. By interpolating the real glottal flow pulse rendering to fundamental frequency F0 and scaling in magnitude according to the energy measure, a pulse train covering a series of separable glottal flow pulses is produced. To regulate the amount of voicing in the excitation, the degree of noise in the excitation is coordinated by employing the phase and magnitude of the spectrum of every pulse according to the harmonic to noise measure at each frequency band. In addition, the spectral tilt of every single pulse is altered affording to the all pole spectrum produced by the HMM. This is accomplished by pass through a filter the pulse train with an adaptive IIR filter which levels the spectrum of the pulse train and spread over the desired spectrum. The lip radiation effect is modeled as a first order differentiation process for voiced excitation. The unvoiced excitation is composed of white noise, whose gain is governed according to the energy measure produced by the HMM system. In order to improve for the over smoothing the vocal tract parameters are enriched and the LSFs are then interpolated and converted to LPC coefficients used for filtering the excitation signal.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method is trained with a prosodic annotated database of 750 phonetically rich sentences. It is spoken by a 35-year-old Tamil male speaker; consist of approximately one and half an hour of speech material. The speech is sampled at 16 kHz. In order to parameterize the spectra of voiced and unvoiced speech 20th order LPC is utilized, and for parameterizing the voice source spectrum 10th order LPC is used. A nine state left-to-right model structure with 7 emitting states are used. Each new feature type was assigned to an individual stream, resulting in a model of 8 streams with a feature order of 171 in total. In the current system, all streams except the fundamental frequency are modelled by a single Gaussian distribution with a diagonal covariance matrix.

In the training part monophony models are trained and then converted to context dependent models. Decision tree state-tying is achieved for each stream for strongly estimate the model parameters. For decision tree clustering, a rich set of contextual features is extracted by a proprietary front-end, ranging from phone level to higher-level phonological features such as word prominence, clause type, and whether the sentence starts a new topic.

To assess the quality of the proposed Tamil TTS system, a Comparison Category Rating (CCR) test is used to assess the excellence of the proposed method in evaluation to natural speech and synthetic speech generated by the system. In this test, the listeners are accessible with a couple of speech samples on each trial. They are examined to evaluate the quality of the synthetic speech compared to the quality of the natural speech on the Comparison Mean Opinion Score (CMOS) scale. Fifteen arbitrarily chosen sentences from held-out data are chosen for creating the test samples for each method. Fifteen Tamil naive spectators (9 men and 6 women) compared a total of 90 speech sample pairs. The ranking is assessed by averaging the scores of the CCR test for each method.

Another test is conducted only for the synthetic sounds produced by the HMM-based TTS systems. Fifteen indiscriminately selected sentences from held-out data (different from the ones used in the CCR test) are utilized for generating the test samples for each method. Fifteen Tamil listeners (9 men and 6 women) compared a total of 44 speech sample pairs. Voice quality testing is performed using subjective test. In subjective tests, human listeners hear and rank the quality of processed voice files according to a certain scale. The most common scale is called Mean Opinion Score (MOS) and is composed of 5 scores of subjective quality, 1-Bad, 2-Poor, 3- Fair, 4-Good, 5-Excellent. The MOS score of a certain TTS system is the average of all the ranks voted by different listeners of the different voice file used in the experiment. The tests are conducted in a laboratory environment with 50 students in the age group of 20-28 years by playing the synthesized Tamil speech signals through headphones.

In this case, the subjects should possess the adequate speech knowledge for accurate assessment of the speech signals and are examined to evaluate the articulacy and spontaneity of the synthesized speech. They have to assess the quality on a 5-point scale for each of the sentences. The mean opinion scores for assessing the intelligibility and naturalness of the synthesized Tamil speech is given in Table-2.

Table-2. Mean opinion score and it is level of confidence of synthesized speech in Tamil language.

Mean opinion score				**Level of confidence** (%)	
TTS with prosody		**TTS without prosody**			
Intelligence	Naturalness	Intelligence	Naturalness	Intelligence	Naturalness
4.4	4.1	3.8	3.2	> 97.5	> 98.5

The MOS scores show the intelligibility of the synthesized Tamil speech is honestly acceptable, whereas the naturalness appears to be little degree of degradation. Naturalness is mostly attributed to distinct perception. It can be enhanced to some degree by integrating the stress and co articulation information along with duration and pitch. The accuracy of the prediction of prosody models can be also analyzed by conducting the listening tests for judging the intelligibility and naturalness on the synthesized speech without incorporating the prosody. In this case, speech samples are the derivative of concatenating the neutral syllables without integrating the prosody. The MOS of the excellence of the synthesized speech without incorporating the prosody have been observed to be low compared to the speech synthesized by combining the prosody. The consequence of the differences in the pairs of the MOS for intelligibility and naturalness is verified using hypothesis testing and the level of confidence is high (>99.5%) for both cases.

The results of the proposed new TTS system exploiting GIF have a significantly better quality than the previously developed HHM based method. Related to natural speech, the quality of the proposed system is clearly inferior. However the prosodic features of the synthetic speech are produced from the HMM, the assessed quality may partially effect from the prosodic discrepancies between the synthetic and natural speech sounds. The proposed system is always preferred over the baseline TTS system. The experimental results show that the proposed system is able to generate natural sounding speech. The development of the presented system continues, and future work will be focused on improving the use and shaping of the natural glottal pulses, and enhancing the use of voice source characteristics obtained by GIF in the synthesis module.

6. CONCLUSIONS

In this proposed work, a HMM-based Tamil text-to-speech system utilizing GIF is described. Fifteen speaker's speech samples and synthesized speech from prototype TTS systems are analyzed. The best speaker who has uniform characteristics of pitch, energy dynamics and speaking rate is selected. Speech corpus has been created from the best speaker is used to build unrestricted TTS. Text corpus has been created from various domains. Optimal unit selection algorithm is used reduce redundancy in the text corpus. The exploitation of GIF in an HMM-based Tamil TTS system is warranted since a large portion of what can be categorized as naturalness in speech appears from the voice source. Thus, exploiting knowledge that describes the working of the real excitation of the human voice construction mechanism might upgrade naturalness of synthetic speech. Thus unrestricted TTS in Tamil language is developed. Based on the subjective quality test results it is conclude that the proposed TTS system producing the synthesized speech with naturalness and good quality.

REFERENCES

[1] Tokuda K., Zen H. and Black A. W. 2002. An HMM-based speech synthesis system applied to English. Proc. IEEE Workshop on Speech Synthesis. pp. 227-230.

[2] Heiga Zen, Keiichi Tokuda and Alan W. Black. 2009. Statistical parametric speech synthesis. Speech Commun. 51(11): 1039-1064.

[3] A. Hunt and A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In ICASSP. pp. 373-376.

[4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata and J. Isogai. 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio, Speech, Lang. Proc. 17(1): 66-83.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura. 1997. Speaker interpolation in HMM-based speech synthesis system. In Eurospeech. pp. 2523-2526.

[6] Gutkin X. Gonzalvo, S. Breuer and P. Taylor. 2010. Quantized HMMs for low footprint text-to-speech synthesis. In Interspeech. pp. 837-840.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King and S. Renals. 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans. Audio, Speech, Lang. Proc. 17(6): 1208-1230.

[8] T. Drugman, B. Bozkurt and T. Dutoit. 2011. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Communication. 53(6): 855-866.

[9] T. Drugman, B. Bozkurt and T. Dutoit. 2012. A comparative study of glottal source estimation techniques. Computer Speech and Language. 26(1): 20-34.

[10] Carlson R., Granstrom B. and Karlsson I. 1990. Experiments with voice modelling in speech synthesis. Speech Commun. 10: 481-489.

[11] Cabral J. P., Renalds S., Richmond K. and Yamagishi J. 2007. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. Sixth ISCA Workshop on Speech Synthesis.

[12] Alku P., Tiitinen H. and Natanen R. 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. Clinical Neurophysiology. 110: 1329-1333.

[13] Raitio T, Suni A, Yamagishi J, Pulakka H, Nurminen J, Vainio M, Alku P. 2011. HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. Audio, Speech, and Language Processing, IEEE Transactions on. 19(1): 153-165.

[14] Sudhakar B., Bensraj R. 2014. An Efficient Sentence-based Sentiment Analysis for Expressive Text-to-speech using Fuzzy Neural Network. 8(3): 378-386.