Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 3– No.6, November 2015 – www.caeaccess.org

Enhanced Evaluation of Sentiment Analysis for Tamil Text-to-Speech Synthesis using Hidden Semi- Markov Model

B. Sudhakar
Assistant Professor
Electrical Engineering Dept
Annamalai University
Chidambaram

R. Bensraj
Assistant Professor
Electrical Engineering Dept
Annamalai University
Chidambaram

ABSTRACT
In recent years, speech synthesis has become an dynamic research area in the field of speech processing due to the usage of automated systems for spoken language interface. This paper address an innovative Tamil Text-to-Speech(TTS)synthesis system utilizing Hidden Semi-Markov Model (HSMM) to analyze sentiments from the speech output. Four different HSMM methods have been proposed to implement the above task. The sentiment-dependent (SD) modeling is the first method, utilize individual models trained for each emotion individually. Sentiment adaptation (SA) modeling is the second method, initially a model is trained using neutral speech, and the adaptation process are implemented to each emotion of the database. The third method is called sentiment-independent (SI) technique, is at first trained utilizing data from all the sentiment of the speech database which is based on an average emotion model. Subsequently, an adaptive model has been constructed for each emotions. The fourth method is called sentiment adaptive training(SAT), the average emotion model is trained with simultaneously normalization of the output and state duration distributions. These training methods are evaluated using a Tamil speech database which consists of four categories of speech, anger, joy, sadness and Disgust was used. To assess and compare the potential of the four approaches in synthesized sentimental speech an emotion recognition rate subjective test was performed. Among the four evaluated methods the sentiment adaptive training method gives the enhanced emotion recognition rates for all emotions.

Keywords
HMM, HSMM,TTS

1. INTRODUCTION
In recent years, TTS synthesis systems have been extensively used in many real time applications for automated speech synthesis around the world. The unit selection corpus-based technique [1] is the conventional approach utilized in the field of speech synthesis. It is primarily based on clustering units of a large speech database according to distance criteria and catches the appropriate ones during runtime based on identical criteria. The naturalness and intelligibility of the speech synthesis system can be improved by using the unit selection corpus-based technique. In this approach, the style characteristics of the synthesized speech track the ones of the recorded speech of the database. This phenomenon will reduce the deviation of speaking styles, sentiments or voice characteristics of the synthetic speech. Because every instant large databases should be recorded to track these variations or styles. The statistical parametric speech synthesis and the hidden Markov model (HMM)-based speech synthesis [2] eliminate the drawbacks of the unit selection speech synthesis approach.

The HMM based techniques are promoted from better adaptability and evidently lesser memory requirement. Initially HMMs are trained utilizing speech databases of natural speech then the synthesized speech is generated through the Mel log spectral approximation (MLSA) filter [10]. Even though it produces synthetic speech of inferior quality, gives some advantage of modeling different speaking styles and emotions with the use of restricted databases.

To enhance the quality of synthetic speech using the restricted speech database the model adaptation algorithms are used. The Maximum Likelihood Linear Regression (MLLR) algorithm [3] or MAP-based (Maximum A Posteriori) modification [4] are the typical examples for model adaptation algorithm. The target speaker is not constrained only to diverse speakers with diverse voice characteristics but also can be characterized by diverse speaking styles or even different emotions. An implementation of diverse speaking styles and emotions plays an important role in the field of HMM based emotional speech synthesis system. A comparative assessment of HSMM (hidden semi-Markov model) training approaches with HMM(Hidden Markov Model)-based speech synthesis for synthesizing speech of various emotions are implemented to eliminate the limitations of HMM technique.

The sentimental speech has been synthesized using four different HSMM training techniques. In the first approach called Sentiment-dependent modeling, the speech data related to each sentiments individually is used in order to train the model producing synthetic speech of this sentiments. The second approach is emotion adaptation modeling, at first a database of neutral speech is utilized for training a model. Subsequently this model is adapted to the target emotion with a MLLR-based adaptation technique using the speech data related to each emotion individually in order to synthesize emotional speech of the specific emotion. The third technique, called emotion-independent modeling, the speech data of all the emotional categories of the database are used to train the average emotion model. In this approach a MLLR-based adaptation technique is used for adapting the model created using all the speech data, to the target emotion. Ultimately a method called emotion adaptive training, based on speaker adaptive training approach [5,6], was implemented. A speech database of emotional speech of Tamil, consisting of four

categories of emotional speech anger, joy, sadness and Disgust along with a second database of neutral speech of Tamil are investigated. To assess the emotion recognition rate subjective test has been taken to estimate the ability of each approach to synthesize the emotional speech.

2. SPEECH SYNTHESIS BASED ON HIDDEN SEMI-MARKOV MODEL

The spectrum, pitch and duration of natural speech are simultaneously modeled in HMM-based parametric speech synthesis system. The spectrum is modeled by continuous probability distribution HMMs in the training phase, the pitch is modeled by multi-space probability distribution HMMs and the state durations by multi-space Gaussian distribution. In order to eliminate the setback of non continuous pitch values in the parts of unvoiced speech multi-space distribution models are used[7]. For clustering separately the distributions of the spectral, pitch and duration parameters context-dependent decisions trees are used. In synthesis segment, a sentence HMM is created by concatenating phoneme HMMs based on the input text. Using a speech parameter generation algorithm[8] speech parameters vector sequences are generated by the concatenated phoneme. At last the synthetic speech is produced through the MLSA filter [9]. An irregularity is raised in this process since state duration models are unambiguously used during the synthesis phase without being incorporated in the training phase. This negative aspect of HMMs can be destroyed with the use of the HSMMs [10]. HSMMs are characterized by their capability to fit in the explicit modeling of state durations not only in the synthesis phase as HMMs do, but also in the training phase of the HSMM-based speech synthesis systems elevate the naturalness of synthetic speech [10].

2.1 Sentiment-Dependent(SD) Method

In this method each sentiments are independently modeled by an acoustic model using only the data associated to this sentiments. A supplementary root node is implemented in the clustering decision tree having as leaves the related decision tree of the respective sentiments. The training and synthesis phases track the relevant procedures which is elaborated in section 2.

2.2 Sentiment Adaptation (SA) Method

In this method a neutral speech database is used to train an initial model in the training stage following the respective procedure. In the adaptation segment, the data related to the target emotion are used to adapt the initial neutral model to the model of the target emotion. An MLLR adaptation [11] is applied for transforming both the output and state duration distributions of the HSMMs. Two categories of regression matrices are produced in this adaptation technique, one for the output distribution and another one for the state duration distribution, so as to maximize the likelihood of the adaptation data [12].The adapted target emotion model is used to generate synthetic speech following the respective procedure in synthesis phase.

2.3 Sentiment-Independent (SI) Method

In this method, using a multi-emotional database an average emotion model is trained, associated to the average voice model [12]. This model is adapted to the target emotion using the respective data. Exclusively, in the training phase, the emotion-dependent models, one for each emotion of the database, apart from the target emotion, are initially individually trained using the multi-emotional speech database. Subsequently using a shared decision tree these context-dependent emotion models are clustered and creating an emotion-independent decision tree. By combining, at each leaf node of the decision tree, Gaussian probability distribution functions of the emotion-dependent models the average emotion model is produced. In order for all the nodes of the decision tree to have data from all the emotions, during the split of a node of the decision tree, only the context related questions which can split the node for all emotion-dependent models, are used. The adaptation of the average model to the target emotion is achieved through a MLLR adaptation trans formation of the output and state duration distributions. In synthesis phase the HMMs generate speech parameters vector sequences through the speech parameter generation algorithm [19] and finally the synthetic speech is generated through the MLSA filter.

2.4 Sentiment Adaptive Training(SAT) Method

The Sentiment adaptive training technique, respectively to the speaker adaptive training normalizes simultaneously the output and state duration distributions of the average emotion model. In SAT, the MLLR adaptation is used as an emotion normalization technique of the average emotion model to reduce the influence of emotion differences and acoustic variability of spectral and pitch parameters [13]. Basically, in the HSMM-based SAT approach the parameter set of HSMM and the set of transformation matrices, for each training emotion in respect to the average emotion model, are estimated simultaneously maximizing the likelihood of the training data. After the average emotion model is trained the adaptation of the model to the target emotion is achieved using MLLR approach [12] and the synthetic speech is generated through the MLSA filter.

3. SPEECH DATABASES

Two Tamil speech databases were used in our experiments, one emotional speech database containing four emotional categories and one database containing only neutral speech .The content of the sentimental speech database was extracted from students, newspapers or were set up by a professional linguist[14]. This database is linguistically and prosodic ally rich, and contains emotional speech from the categories: anger, joy, sadness and Disgust which are considered as the four emotions as well as neutral speech. The database consisted of 70 utterances, which were pronounced several times with different emotional charge The length of the utterances was ranging from a single word, a phrase, short and long sentence or even a sequence of sentences of fluent speech. The context of all sentences was emotionally neutral, meaning that it did not convey any emotional charge through lexical, syntactic or semantic means. The entire database consisted of 5000 words. All utterances were uttered by a professional, female singers , speaking Tamil[15]. All recording sessions were held in the chamber of a professional singers recording room.

3.1 Listening Test Performance

A listening test was implemented to measure the emotions of the database from the recorded speech database. Five listeners, of dissimilar ages with no particular knowledge in speech synthesis, were asked to recognize the emotion that characterized each recorded utterance . Five sentences were selected with all the relevant emotions and played randomly to each listener. In the first portion of the test, a free response was

given by the listener classification each recording with whatever emotion found appropriate. Subsequently the second portion test follows forced response test, the listener was classifying each recording to one of the four emotional categories included in the database (anger, joy, sadness and Disgust). In Table 1the results of Performance measure on different sentiments of speech database are given.

Table1:Performance measure on different sentiments from the database

Sentiment Types	Free Response Test (%)	Forced Response Test (%)
Anger	9.6	97.3
Joy	85.1	90.4
Sadness	94.4	95.0
Disgust	65.0	71.5

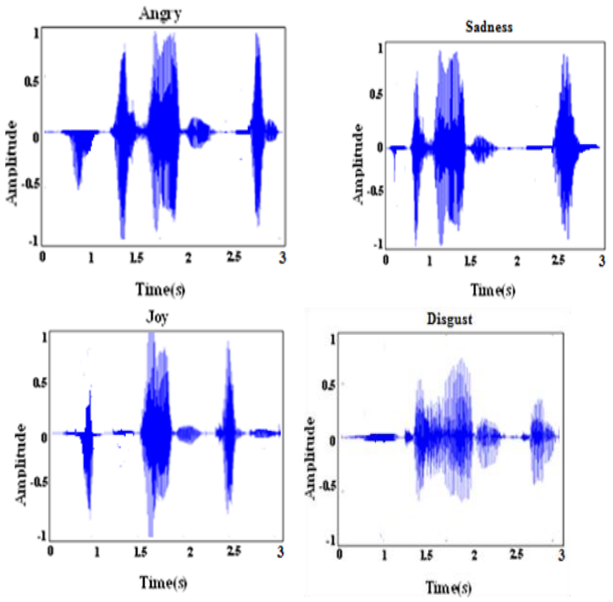

Fig 1: Forced response test of different sentiments

3.2 Implementation Procedure of HSMM TTS System

In our experiments, the speech signals of both databases were down-sampled to the frequency of 18 kHz and a phone set of 38 phones was adopted for building the HSMM-based speech synthesis systems. We used 5-state left-to-right with no-skip HSMMs. The parameters were extracted using a 30 msec Hamming-windowed frame length with a 10 msec frame shift. The feature vector consisted of 30 Mel-frequency cepstral coefficients (MFCCs), including zeroth coefficient, and logarithm of fundamental frequency (logF0). Moreover, both dynamic (delta) and acceleration (delta-delta) coefficients were used both for spectrum (MFCC) and pith (logF0) representation.

In the case of SD method each model was trained using 70 sentences of the respective emotion and the evaluation of them was done synthesizing the rest 10 sentences. Concerning the SA modeling, 1100 sentences of neutral speech database were used for training the initial neutral model. For the adaptation of the model, the same 70 sentences with the ones used in the SD modeling were used for each emotion and the rest ones (10 sentences) for the evaluation of the approaches. In the case of SI method, the 1100 sentences of neutral selected from neutral speech database along with the same 70 sentences mentioned above for each emotion apart from the ones of the target emotion were used for training the SI model. The 70 sentences of the target emotion were used for the adaptation of the average emotion model created by the training phase of the SI modeling and the rest ones (10 sentences) of the target emotion were used for the evaluation of the model. The same sets of sentences were also used for the average emotion model created by the SAT approach.

4. RESULTS AND DISCUSSION

Table 2:Performance analysis of different sentiments Using HSMM Method

Sentiment Types	SD %	SA %	SI %	SAT %
Anger	96.7	97.1	97.8	98.3
Joy	80.3	81.5	82.3	84.5
Sadness	95.2	95.8	96.5	97.2
Disgust	65.2	66.4	68.1	70.8

The efficiency of the different training approaches are estimated through the performance analysis of different sentiments using HSMM Method. Five males and three females have been selected for the emotion classification test, were asked to categorize the synthesized utterances to the emotional categories. Each candidates are presented with ten synthesized sentences for each one of the four training approaches (SD, SA, SI, SAT) and for each one of the four emotions (anger, joy, sadness, and Disgust). The results of the subjective test are illustrated in Table 2. The results prove that the SAT method produced the most excellent emotion recognition rates followed by the SI ,SD and the SA models. The SD method produced the lowest performance in the emotion classification performance analysis evaluation test. The performance enhancement and degradation of each methods are based on the basic principles of each method and the available training data. The SD method produce the lowest score for each emotional category, due to the scarcity of available training data. In the SA method, the existing data of emotional speech are used only for adapting the neutral model to the target emotion. In the SI method the average emotion model is fabricated based on the clustering of the context-dependent emotion models to a shared decision tree, training more robust models, better than the SA models. Moreover, in the SAT method, the MLLR adaptation which is used as an emotion normalization technique, managed to decrease the influence of emotion differences in respect to the SA method, construction more robust models and achieving the greatest performance throughout all the emotional categories.

5. CONCLUSION

From this proposed work a comparative assessment on HSMM training approach with HMM-based TTS synthesis approach for producing emotional speech is presented. In the training phase HMM do not support the explicit modeling of state durations, instead the HSMM were used for implementation. A Tamil speech database of sentimental speech, consisting of four categories of sentimental speech anger, joy, sadness, and Disgust was used. Four divers training approaches of HSMM-

based synthesis of emotional speech were analyzed. A performance test was implemented for assess the effectiveness of the training approaches to produce TTS synthetic speech of precise emotions. The sentiment-dependent modeling approach achieved the worst performance. The experimental results proves that the sentimental adaptive training approach achieved the greatest sentimental recognition rates followed by the remaining methods.

6. REFERENCES

[1]. Donovan R, Woodland P. A hidden Markov-model based trainable speech synthesizer [J]. Computer Speech and Language, 1999, 13(3): 223-241.

[2]. Masuko T, Tokuda K, Kobayashi T, Imai S. Speech synthesis using HMMs with dynamic features [C]. In Proceedings of ICASSP, 1996, 389-392.

[3]. Tamura M, Masuko T, Tokuda K, Kobayashi T. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR [C]. In Proceedings of ICASSP, 2001, 805-808.

[4]. Gauvain J, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains [J]. IEEE Trans. Speech Audio Processing, 1994, 2(2):291-298.

[5]. Anastasakos T, McDonough J, Schwartz R, Makhoul J. A compact model for speaker-adaptive training [C]. In Proceedings of ICSLP, 1996, 1137-1140.

[6]. Yamagishi J, Kobayashi T. Adaptive training for hidden semi-Markov model [C]. In Proceedings of ICASSP, 2005, 365-368.

[7]. Tokuda K, Masuko T, Miyazaki N, Kobayashi T. Hidden markov models based on multi-space probability distribution for pitch pattern modeling [C]. In Proceedings of ICASSP, 1999, 229-232.

[8]. Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T. Speech parameter generation algorithm for HMM-based speech synthesis [C]. In Proceedings of ICASSP, 2000, 1315-1318.

[9]. Fukada T, Tokuda K, Kobayashi T, Imai S. An adaptive algorithm for mel-cepstral analysis of speech [C]. In Proceedings of ICASSP, 1992, 137-140.

[10]. Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T. Hidden semi-Markov model based speech synthesis [C]. In Proceedings of ICSLP, 2004, 1180-1185.

[11]. Yamagishi J, Masuko T, Kobayashi T. MLLR adaptation for hidden semi-Markov model based speech synthesis [C]. In Proceedings of ICSLP, 2004, 1213-1216.

[12]. Yamagishi J, Tamura M, Masuko T, Tokuda K, Kobayashi T. A context clustering technique for average voice model in HMM-based speech synthesis [C]. In Proceedings of ICSLP, 2002, 133-136.

[13]. Yamagishi J, Kobayashi T. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training [J]. IEICE Trans. Inf. & Syst, 2007, E90-D(2):533-543.

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

[15]. Sudhakar.BandBensraj.R.2015.AnexpressiveHMM-BasedText-To-SpeechSynthesis System Utilizing Glottal Inverse filtering for Tamil Language. ARPN Journal of Engineering and Applied Sciences.10(6):2400-2404.