PRESERVATION OF SPEAKER IDENTITY IN HMM BASED SPEECH SYNTHESIZERS
Speech Signal Processing

M. Niranjana Priyadarshini, S. Sathya Devi, M. Divya, D. Gayathri

Dept. of ECE, PSNACET, Dindigul, India.
priyadarshinim89@gmail.com, sathyadevi.s@gmail.com, deivagayathri@gmail.com, divyapsna87@gmail.com

ABSTRACT

The purpose of speech synthesizer is to convert the input text to speech. Most common synthesizers are USS and HTS synthesizer. The speech synthesized by the HMM based system is found to be more intelligible than that synthesized by the USS system due to the elimination of sonic glitches and also the memory requirement of HMM based system is less around 5MB as against 500MB for an USS system. Hence HMM based synthesizer is efficient and economical. But in HMM synthesizer, buzziness is detected which prevents the preservation of speaker's identity and decreases the intelligibility and pleasantness. The speaker's identity preservation depends on two parameters namely speech rate and number of states used for modelling. When the input speech rate is slow, the generated speech seems to be noisy as the number of formants per window is not sufficient. Hence, the importance of choosing an appropriate speech rate in text to speech synthesis systems is analyzed. Two 3-hour speech corpora – one with a slow speech rate of 8 phonemes/sec and the other with a fast speech rate of 11 phonemes/sec are developed for the South Indian language, Tamil. The effect of speech rate on the synthetic speech produced by four systems, namely, phoneme-based and consonant-vowel (CV) unit - based, unit selection synthesis (USS) and hidden Markov model (HMM) – based synthesis systems, are analyzed. Based on a subjective evaluation using the mean opinion score (MOS), it is observed that three systems perform better when trained with the fast speech corpus, with MOS ranging between 2.78 and 3.18, except in the case of CV based USS. Further, it is observed that the phoneme-based systems perform better than the CV-based systems, with three hours of data.

Index Terms— Speech synthesis, speech rate, HTS, USS, CV units.

I. INTRODUCTION

A speech synthesizer converts a given input text into the corresponding speech waveform. Among various speech synthesis approaches, the unit selection approach and hidden Markov model (HMM) based approach are most successful. Unit selection synthesizers (USS) concatenate the pre- recorded speech units that best match the given input text and synthesize speech, whereas, in hidden Markov model based speech synthesis system (HTS) instead of storing the speech waveform, the features extracted from the recorded speech are stored as models. These models are concatenated based on the given text to form a sentence HMM. Then the spectral and excitation features extracted from the sentence HMM are used to synthesize speech [1]. With the use of HTS, synthetic speech with increased perceptual quality can be synthesized with smaller corpora [2].

The quality of synthetic speech is influenced by 2 factors. They are the speech rate of the training data [3], [4] and the length of the synthesis unit. The effect of speech rate is concentrated in the current work expecting that, with the use of a fast training data, the perceptual quality will increase [3, 4]. In addition, the effect of using longer unit such as the consonant-vowel (CV) unit is also analyzed. The use of CV units is expected to increase the quality due to the conservation of various contextual factors within a single unit and the capture of co-articulation [5,8]. [5,8] also discuss various longer units that are capable of capturing the contextual variations.

The speech synthesizers incur the limitation that with the use of longer units, the size of the database increases [8]. It also suffers from a necessity that with the use of longer units, there should be sufficient number of occurrences for every synthesis unit in the data base, as some of the synthesis units may have rare occurrences or may not occur at all [9]. When such rarely occurring or missing units are needed during synthesis, then the system suffers from a loss in quality. Such issues regarding missing units are taken care by the tree based clustering in HTS, which provides similar acoustic models for unseen units.

In the current work, the effect of speech rate of the training data, on the quality of synthetic speech is analyzed using 2-3-hour Tamil speech corpora. The analysis is performed in four systems, namely 1) Phoneme based USS 2) CV based USS 3) HTS with phoneme as a basic unit along with 2 left and right contexts (penta-phoneme) and 4) HTS with CV as a basic unit with 2 left and right contexts (penta CV). The performance of the developed systems is analyzed by a subjective evaluation using the mean opinion score (MOS).

The paper is organized as follows. In Section II the steps involved in speech corpora development is explained. Section III gives an introduction to the unit selection TTS (text to speech synthesis) system and Section IV describes in brief about the HMM based speech synthesizer. Finally, Section V deals with the performance analysis of the systems developed and Section VI concludes the paper.

II. SPEECH CORPORA AND DATA PREPARATION

For building the 3hour Tamil speech corpora, initially the text data void of colloquial words is collected and normalized by removing punctuations. The normalized text data consists of 1800 sentences from the Tamil novel, "Parthiban Kanavu" [10]. The sentences were recorded at 2 different rates. The fast speech corpus was recorded with all 1800 sentences, whereas only 1100 were used for developing the slow corpus. The speech data is recorded from a native female speaker, using a mixer and a handheld micro-phone at a sampling rate of 16 kHz, in a laboratory environment. The 2 corpora are developed at speech rates of 8 phonemes/sec

and 1 phonemes/sec, where the normal speech rate varies between 10-17 phonemes/sec [11]. Drastic variations in voice characteristics of the speaker are avoided. Moreover, the pleasantness, clarity, loudness, quality, speech rate and pronunciations are taken into consideration. The text data is transliterated based onITRANS

(Indian languages transliteration) standard for Tamil [12]. For synthesis, both the recorded speech files and time aligned phonetic transcriptions are needed, where the label files help in matching the text with the speech waveform and provides information about the occurrence of the speech units. Hence, the recorded speech needs to be segmented.

A phoneme level segmentation is performed manually for 5 minutes of data. The 5 minutes of data consists of 13 phonetically balanced sentences. Labelling is carried out with the help of visual representations such as the waveforms and spectrograms. For efficient manual segmentation of the speech data a prior knowledge about the characteristics of different phonemes is needed [13, 14]. Each phoneme has its unique spectral characteristic. Vowels have a well-defined spectrogram and occur for longer duration. Unvoiced consonants occur at the higher frequency region and have very short durations whereas voiced consonants occur at the lower frequency region but are not well structured as that of vowel [13, 14]. Accurate segmentation is essential as it influences the quality of synthesis. The manually segmented data is later used for training models which in turn are used to perform the forced Viterbi alignment [15, 17], procedure to obtain phoneme-level label files for both the corpora.

These phoneme level label files are then modified to obtain CV label files, where a consonant is concatenated with its succeeding vowel, if not succeeded by a vowel it is left unchanged as a single phoneme. For example, consider the word “thIram”, for phoneme based system it is segmented as /th/ /I/ /r/ /a/ /m but for CV based system it is yielded as /thI/ /ra/ /m/.

III. UNIT SELECTION TTS SYSTEM

In the Fig. 1 (redrawn from [18]), different stages of a unit selection TTS system are portrayed. The text input is given to the text pre-processing and prosodic phrasing units, which divides the text based on punctuations. The pronunciation generation block generates the basic units based on the lexicon and the letter to sound rules (LTS), where the lexicon is a list of all speech units and the LTS splits each word based on the unit that is to be used for synthesis [18]. The classify-cation and regression trees (CART) database has binary hierarchical decision trees, which checks the requisite feature for the text and reaches the leaf node containing the best matching synthesis unit or cluster of best matching units with similar properties [18].

The offline database preparation depicts the data preparation and its segmentation for training the system. The most important block is the unit selection algorithm which provides the best sequence of speech units by estimating the target cost [19] (matching estimate) and concatenation cost [19] (estimate for combination of units). Finally, the waveform of these speech units is concatenated to produce the synthetic speech.

Festival framework offers a unit selection approach for synthesizing speech. It uses different phonetic and prosodic contexts that are available in the database for synthesizing speech with the required characteristics. In the current work, the unit selection synthesis performed is an unrestricted domain speech synthesis.

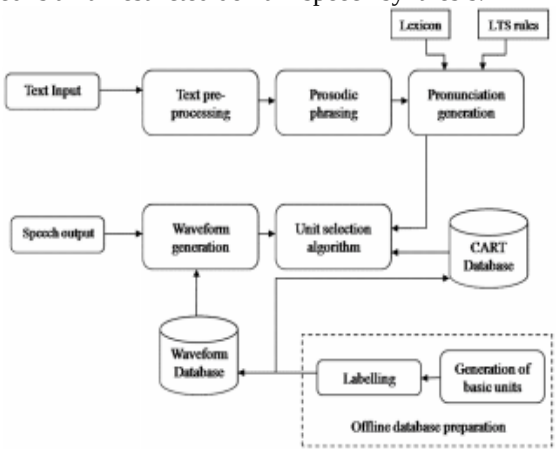

Fig. 1. Unit Selection TTS system (redrawn from [18]).

Unit selection based speech synthesizers using phoneme and CV units are developed with the slow and fast rate corpora. The training of the synthesis system requires the recorded and segmented data along with the text data. The slow speech corpus contains 39 phonemes for the phoneme based system and 12 vowels, 25 consonants and 225 CV units for the CV based system. Similarly, the fast speech corpus contains 39 phonemes for the phoneme based system and 12 vowels, 25 consonants and 240 CV units for the CV based system.

The steps to be followed for building a new voice in Festival framework are as follows:

- Phone set is created which defines the features such as vowel length, height and frontness, lip rounding, consonant type, place of articulation and consonant voicing, for each sound unit. It consists only of phonemes in the case of a phoneme based system. In the case of a CV based system it also consists of the features of CV units. Thus, the phone set is formed by enumerating all the possible units and defining its features.
- The units in the training database are clustered based on the textual and acoustic features using Wagon tool [20, 21].
- An utterance provides the basic structure for synthesis. It defines the phonetic information, duration, prosodic information and pitch for each speech unit of the text to be synthesized or text used to build the database. It also gives the position of the speech unit in the word it occurs and holds phonetic information about the units adjacent to it.

Thus, after building the system, when an input text is given, it is split in to the required speech units and the best matching units are found from the database and concatenated together and speech is synthesized.

IV. HMM BASED SPEECH SYNTHESIS SYSTEM

The HTS is a statistical parametric speech synthesis system. It has less memory requirements. It consists of

two main phases as shown in Fig. 2 (redrawn from [2]), the training and the synthesis phase.

Training phase of HTS is similar to that of speech recognition, but the uniqueness is that it extracts the excitation parameters and durations to restore the temporal structure of speech along with the spectral parameters [2]. In the synthesis part, the given input text is converted in to a context based label sequence and then the sentence HMM is constructed by concatenating the appropriate context dependent HMMs [2]. The time aligned transcriptions and the final utterances containing information about the contextual features (53 contextual features), which is used in building the systems in Festival framework is used in HTS to derive the context dependent label files.

The parameters namely duration parameters, excitation parameters (log Fo) and the spectral parameters are determined using a speech parameter generation algorithm, in such a way that the output probability for HMM is maximized. Excitation parameters has 3 dimensions including its dynamic features, whereas Mel-generalized coefficients (MGC) is 105-dimensional consisting of 35 static, 35 first derivative and 35 second derivative features. based clustering is out the of -set to generate context dependent models. The question set contains the questions (57 questions) about the features that help in sorting the various speech units in to clusters with similar properties. Finally, the parameters are used to drive a speech synthesis filter and the synthetic speech is produced.

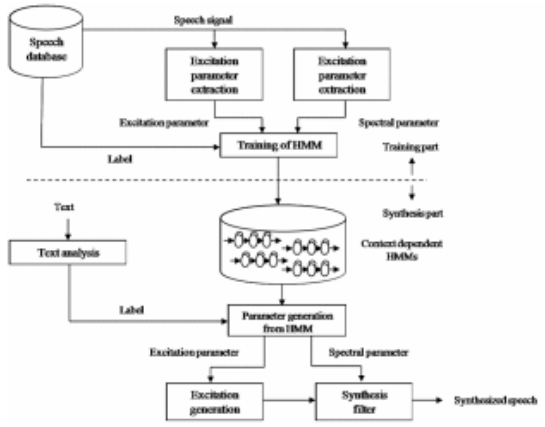

Fig. 2. HMM based Speech Synthesis System (redrawn from [2]).

HMM based speech synthesizers using phoneme and CV units are developed with the slow and fast rate corpora. The synthesis requires the utterance data from Festival framework and speech data. The number of phoneme and CV units for the fast and slow corpora is given in Section III. The other advantages of HTS are that it facilitates speaker adaptation and modification of voice characteristics by means of modifying the HMM parameters as discussed in [2, 22, 23].

V. Evaluation and Analysis:

The performance of the systems is evaluated subjectively using a listening test. The listening test is conducted with 8 normal listeners and it is carried out in a laboratory environment using a loud speaker. Synthesized speech is played simultaneously to all the listeners. Totally, 8 systems are subjected for evaluation. 1) Phoneme based USS using slow corpus, 2) CV based USS using slow corpus, 3) HTS with phoneme as a basic unit along with 2 left and right contexts (pentaphoneme) using slow corpus, 4) HTS with CV as a basic unit with 2 left and right contexts (penta CV) using slow corpus, 5) Phoneme based USS using fast corpus, 6) CV based USS using fast corpus, 7) HTS with phoneme as a basic unit along with 2 left and right contexts (pentaphoneme) using fast corpus and 8) HTS with CV as a basic unit with 2 left and right contexts (pentaCV) using fast corpus.

TABLE I: Rating Criterion

Score	Quality and impairment
5	Excellent, perceptible
4	Good, perceptible but annoying
3	Fair, slightly annoying
2	Poor, annoying
1	Bad, very annoying

The evaluation is carried out by playing the speech synthesized by each system at considerable intervals to avoid the perceptual influence of one system over another when played sequentially. 15 sentences synthesized by each system are used for evaluation. The perceptual quality of the synthesized speech is rated on a scale of 1 to 5, where a rating of 5 depicts a very high perceptual quality and 1 denotes the bad and annoying experience as shown in Table I [24]. After scores are obtained from all the listeners, the mean is calculated. Table II gives the mean opinion score (MOS) obtained for all the 8 systems. The MOS in Table II reveals that the systems trained with a fast speech corpus perform better than those trained with the slow speech corpus, except in the case of CV based USS.

TABLE II: Mean Opinion Score

Speech rate and Synthesis Systems / Synthesis Units	Slow speech corpus		Fast speech corpus	
	USS	HTS	USS	HTS
Phoneme based	2.75	3.16	2.78	3.1
CV based	2.72	2.88	2.22	2.9

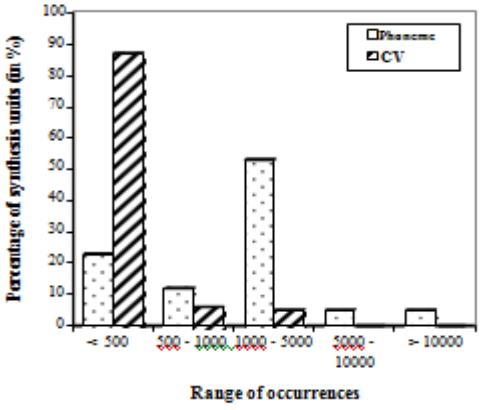

Fig. 3. Distribution of no. of examples for the synthesis units.

It is found that the phoneme based USS using fast corpus has obtained a MOS of 2.78. Similarly, the phoneme and CV based HTS using fast corpus has obtained a MOS of 3.18 and 2.92 respectively. The CV based USS built using fast corpus obtained only a MOS of 2.22 which is less than the CV-USS using slow corpus. The loss in quality is because of the loss of clarity in the fast corpus. Moreover, due to the fast movement of the articulators, the interference due to the preceding or succeeding articulatory motion occurs. These interferences are also captured by the CV units while capturing the co- articulation.

The MOS of the phoneme and CV based USS using slow corpus is 2.75 and 2.72 respectively. Similarly, the MOS of the phoneme and CV based HTS using slow corpus is 3.16 and 2.88 respectively. It is also seen that in both USS and HTS the CV based system had MOS less than the phoneme based system when using both of the corpora. The inferior performance of the CV based system is because of the less number of examples available for the CV units in the speech corpora. Fig. 3 represents the distribution of the examples for the synthesis units in the phoneme and CV based systems using slow corpus. The distribution is calculated using five ranges of occurrence (< 500, 500 -1000, 1000 – 5000, 5000 – 10000, > 10000) plotted against the percentage of synthesis units that occurs in the corresponding range of occurrence. It is well seen that the HTS outperforms the USS in terms of intelligibility, because of the well captured co-articulation with the use of contexts and the absence of sonic-glitches as in the case of USS.

VI. CONCLUSION

In the current work, the phoneme and CV based USS and HTS are developed using slow and fast corpus and the synthesized speech are analyzed. The improvement in the perceptual quality is evaluated by conducting the listening tests. The increase in the perceptual quality and naturalness, with the increase in speech rate is clearly revealed from the analysis for both USS and HTS, except when CV unit is used in USS. With the use of longer unit such as the CV unit more examples are needed and hence a larger database is needed. The improvement in the perceptual quality with the use of CV unit is well seen especially during synthesis of longer words, in the CV based systems. The comparatively better performance is given by the HTS while using the same speech corpus due to the absence of sonic-glitches at the concatenation points as in the case of USS. But yet HTS suffers from limitations such as loss of speaker identity and buzziness in the synthesized speech. Thus, for a successful USS or HTS, that produces synthetic speech with naturalness, the speech rate should be between 10 – 17 phonemes/ second.

REFERENCES

[1] Youcef Tabet and Mohamed Boughazi, Speech synthesis techniques. A Survey, Proc. IEEE 7th International Workshop on Systems, Signal processing and their applications 3: 1712 – 1715 (2011).

[2] Keiichi Tokuda, Heiga Zen and Alan W. Black, An HMM-based speech synthesis system applied to English, Proc. IEEE Workshop on Speech Synthesis Pp. 227 – 230 (2002).

[3] Donata Moers, Petra Wagner and Stefan Breuer, Assessing the adequate treatment of fast speech in unit selection speech synthesis systems for the visually impaired, Proc. 6th ISCA Tutorial and Research Workshop on Speech Synthesis Pp. 282 – 287 (2007).

[4] Donata Moers, Petra Wagner, Bernd Mobius, Filip Mullers and Igor Jauk, Integrating a fast speech corpus in unit selection speech synthesis: Experiments on perception, segmentation and duration prediction, Proc. Speech Prosody Pp. P2A – 28 (2010).

[5] Shirbahadurkar. S.D. and Bormane. D.S., Speech synthesizer using concatenative synthesis strategy for Marathi language (Spoken in Maharashtra, India). International Journal of Recent Trends in Engineering 2(4): 181 – 185 (2009).

[6] Yuki Yoshida, Shin'ya Nakajima, Kazuo Hakoda and Tomohisa Hirokawa, A new method of generating speech synthesis units based on phonological knowledge and clustering technique, Proc. IEEE 4th Int. Conf. Spoken Language, PA 3: 1712 – 1715 (1996).

[7] Yoshinori Sagisaka, Speech synthesis by rule using an optimal selection of non-uniform synthesis units, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, NY 1: 679 – 682 (1998).

[8] Mats Blomberg and Kjell Elenius, Creation of unseen triphones from diphones and monophones using a speech production approach, Proc. IEEE 4th Int. Conf. Spoken Language, PA 4: 2316 – 2319 (1996).

[9] Simon King, An introduction to statistical parametric speech synthesis, Indian Academy of Sciences, 36(5): 837 – 852 (2011).

[10] Parthiban Kanavu [online]. (accessed on 4th August (2012) Available: http://www.projectmadurai. org/pm_etexts/utf8/pmuni0214.html
Tom Brondsted and Jens Printz Madsen, Analysis of speaking rate variations in stress-timed languages, Proc. 5th European Conference on Speech Communication and Technology, Rhodes Pp. 481 – 484 (1997). Avinash Chopde. ITRANS Tamil Table [online]. (accessed on 7th December 2012) Available: http://www.aczoom.com/itrans/html/tamil/node5.html, George Boeree. C, [2005]. Phonetics [online]. (accessed on 6th October 2012) Available:
http://webspace.ship.edu/cgboer/phonetics.html

[11] Quatieri. T.F., Production and Classification of Speech, in Discrete-Time Speech Signal Processing: Principles and Practice, Prentice-Hall Inc., Ch. 3 (2002).

[12] Steve Young, Julian Odell, Dave Allason, Phil Wood land, The HTK book (for version 2.1), Cambridge University Engineering Department, pp. 199 – 209 (1997).

Joseph Picone. Automatic speech Recognition [online]. (accessed on 5th October 2012) Available: http://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/current/section_04/s04_04_p01.html

[13] Ramani Boothalingam, V. Sherlin Solomi, Anushiya Rachel Gladston, S. Lilly Christina, P. Vijayalakshmi, Nagarajan Thangavelu, Hema. A. Murthy, Development and evaluation of unit and HMM- based speech synthesis systems for Tamil NCC (2013).

[14] Samuel Thomas, Natural sounding Text to Speech Synthesis, M.S thesis, Dept. of Computer Science and Eng., IIT Madras (2007).

[15] Alan W Black, Heiga Zen, Keiichi Tokuda, Statistical parametric speech synthesis, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, HI 4: IV-1229 – IV-1232 (2007).

Simon King, Alan W. Black, Paul Taylor, Richard Caley, Robert Clark. Edinbugh Speech Tools Library [online]. (accessed on 30th October 2012) available:http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0

[21] Black, A.W., P. Taylor and R. Caley, The Festival Speech Synthesis System," System documentation (1999).

[22] Divya Bansal, Ankita Goel, Khushneet Jindal, Punjabi speech synthesis system using HTK, International Journal of Information Sciences and Techniques 2(4): 57 – 69 (2012).

[23] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, AlanW. Black, Keiichi Tokuda, The HMM-based Speech Synthesis System (HTS) Version 2.0, Proc. IEEE 6th ISCA Workshop on Speech Synthesis, pp. 294 – 299 (2007).

[24] Wikipedia. Mean Opinion Score [online]. (accessed on 30th, November 2012) Available: http://en.wikipedia.org/wiki/Mean_opinion_score