The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Post-Processing Using Speech Enhancement Techniques for Unit Selection and Hidden Markov Model-based Low Resource Language Marathi Text-to-Speech System

Sangramsing Kayte and Monica Mundada

Speech Research Lab, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

bsangramsing@gmail.com, monicamundada5@gmail.com

Abstract

A speech signal captured by a distant microphone is generally contaminated by background noise, which severely degrades the audible quality and intelligibility of the observed signal. To resolve this issue, speech enhancement has been intensively studied. In this paper, we consider a text-informed speech enhancement, where the enhancement process is guided by the corresponding text information, i.e. a correct transcription of the target utterance. The proposed Unit Selection Synthesis (USS) and Hidden Markov Models (HMM)-based framework are motivated by the recent success in Text-to-Speech (TTS) research. The primary aim of the study is to improve the quality of speech after synthesizing voice employing USS and HMM methods for building low resource Marathi TTS using speech enhancement techniques. Taking advantage of the nature of USS and HMM that allows us to utilize disparate features in an inference stage, the proposed method infers the clean speech features by jointly using the observed signal and widely-used TTS features derived from the corresponding text. In this paper, we first introduce the background and the details of the proposed method for low resource Marathi language. Then, we show how the text information can be naturally integrated into speech enhancement by utilizing USS and HMM and improve the synthesis speech enhancement performance. The spectral subtraction method is used to remove the noise from synthesized speech and improve the quality. The spectral parameters of both the methods shows the progress in the enhanced speech.

Index Terms: Text-to-Speech, Unit Selection Synthesis, Hidden Markov Models, Speech enhancement, Speech corpus, Statistical Parametric Speech Synthesis

1. Introduction

The growing the demand of smart phones making the remarkable progress in development of human-machine-communication, e.g. speech recognition and speech synthesis [1]. The applications related to communication are developed to ask questions and get them answered directly by speech on both sides [2]. Text to Speech (TTS) is a system in which sequence of words are taken as input and converts them to speech. In conversion process of speech synthesis method, in Marathi language vowels and consonants are considerably important [1]. Each syllable is a combination of consonants and vowels. Marathi is a low resource magnanimous language with lots of words directly taken from the Sanskrit language. The online speech corpus for Marathi language is very less as compared to other languages like English, Chinese, thus makes it falls into category of low level languages. The Unit Selection Synthesis (USS) and Hidden Markov Models (HMM)-based TTS system has been popularly studied because of its reasonable quality and easy implementation. Here the Marathi based TTS system is developed using these techniques [3]. The output speech synthesized using these techniques suffers from the unnaturalness of human voice. Speech enhancement applied to synthesize voice meets the parameters of TTS system i.e. understandbility and naturalness [4] [5]. The important methods for enhancing speech are the removal of background noise [6], echo suppression and the process of artificially bringing certain frequencies into the speech signal [6]. It is well known that the naturalness of synthesized speech is improved by adopting efficient excitation and spectrum models[7]. At initial phase, Gamma tone filtering method is applied as sub banding for synthesized voice [8]. Spectral subtraction is used to enhance the synthesized speech for USS and HMM-based TTS system [9]. So, the spectral parameters are studied further of original and enhanced speech with these filters and results in better performances [10].

This paper is organized as follows: the brief description of USS and HMM-based method for building of TTS system is given in Section II. The speech enhancement method for USS and HMM are provided in Section III. The approaches for speech enhancement in the HMM and USS is given in section IV. The experimental analysis is given in Section V and Section VI presents the conclusion.

2. Text-to-Speech Synthesis

TTS synthesis technique defines the generation of artificial speech with a scope to generate intelligible and natural-sounding speech for any given input text. There are the number of techniques and methods developed so far, but the two main TTS techniques include the USS and Statistical Parametric Speech Synthesis (SPSS) approach using HMM for speech synthesis. The next sub-section discusses the working of both the methods in detail [1].

2.1. Unit Selection Synthesis

The USS synthesis always intended to have a well-organized and huge speech corpus. In this, speech database contains the units of specified speech, which is sensibly proposed to have a large coverage of all prosodic and phonetic variants of each unit [2]. In the corpus, each described speech unit has a number of achievable variants, which are suitable to perform in different phonetic and prosodic contexts [4]. The huge speech corpus is analysed offline and all the respective calculated features are stored in a unit database. In the database, each instance of a unit is described by a vector of features [5]. Each feature may have

10.21437/SLTU.2018-20

a discrete or continuous value. The features explain the unit itself and the related context of the given input text. The features of the unit itself are helpful for choosing the correct unit that assembles the segmental requirement, on the other hand the features of context are used for picking the contextually best unit, this will result in minimizing the discontinuity among the selected units for generating the synthetic speech. [6]. The development of USS-based TTS system can be found in the following modules.

Text Processing: In this stage, input text is cleaned and several non-standard symbols, words, punctuations marks, abbreviations, tags, smileys, etc.

Phonetic Analysis: The phonetic analysis refers to the generation of a sequence of speech sound units from the text. This can make use of a dictionary to the mapping from orthography to pronunciation is not always straightforward. The language rules, i.e. Letter-to-Sound (LTS) rules can also be used as done in Indian languages [11].

Prosodic Analysis: Intonation and duration modeling for the given text is covered in this stage.

Speech Generation: The speech generation can be done using rule-based concatenative approach. In cluster unit-selection, speech sound units can be clustered based on acoustic distance. Each of the speech sound units in the data is clustered into similar acoustic groups based on the information at synthesis time, i.e., phonetic context, prosodic features and other higher-level features such as the position of a word, stress, accent, etc. To cluster units, an acoustic measure is defined from acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), fundamental frequency (F0), and delta cepstrum [12].

2.2. Hidden Markov Model

In the HMM-based speech synthesis system [13], rhythm and tempo are controlled by the state duration probabilities modelled by the single Gaussian distributions [14]. They are estimated from statistical variables obtained in the last iteration of the forward-backward algorithm [15], and then clustered by a decision tree-based context clustering algorithm [16] they are not re-estimated in the Baum-Welch iteration. In the synthesis stage, we construct a sentence HMM and determine the state durations maximizing their probabilities [17]. Then a speech parameter vector sequence is generated. However, there is an inconsistency, although parameters of HMM are re-estimated without explicit state duration probability distributions, speech parameter vector sequence is generated from the HMM with explicit state duration probability distributions [18]. This inconsistency might degrade the quality of outputs. This HMM synthesis is divided into training part and synthesis part [19] are given as follows.

Training part: The spectral and excitation parameters are extracted from speech database. The MFCC along with their dynamic features are generally taken as spectral (i.e., vocal tract system) parameters and log (F0) and its dynamic features are taken as excitation (i.e., speech source) parameters. These features are modelled by context-dependent HMM in a unified framework

Synthesis part: Given a test sentence which is to be synthesized, its corresponding utterance is converted to context-dependent phoneme sequence [20]. According to the phoneme sequence, utterance HMM is constructed by concatenating context-dependent HMMs followed by determination of state duration of HMM [21]. Thereafter, using speech parameter generation algorithm, spectrum and excitation parameters are generated. Finally, the speech waveform is generated using Mel Log Spectrum Approximation (MLSA) filter [22].

3. Speech Enhancement Techniques

The speech enhancement methods vary depending upon the kind of degradation. The speech enhancement techniques are classified into two basic categories: Single channel and Multiple channels based on speech recorded using single microphone or multiple microphone sources respectively [23]. The frequently and conventional methods are transforming domain methods [24]. They transfer the time domain signal into other domain using different transforms and involve various filtering methods to suppress noise and then inverse transform the filtered signal into the time domain. The method of spectral subtraction (SS) is used for re-synthesizing speech. Most speech enhancement methods can be described [25] in three steps show the Figure 1:

- Decomposing the speech signal into a transformed domain
- Estimating the clean channel signals in the transformed domain
- Synthesizing the speech from the estimated channel signals

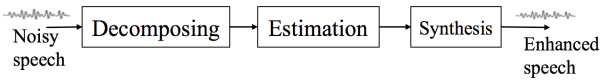

Figure 1: *General block diagram of speech enhancement*

3.1. Spectral Subtraction

Spectral subtraction is a widely used algorithm in acoustic noise reduction, mainly because of its simplicity of implementation. This falls into the category of transform domain methods [26]. In this method, it initially transfers the time domain signal into other domain using different transforms and with help of some filters, it suppresses noise and then inverses transform the filtered signal into the time domain. Spectral subtraction (SS) is used to remove an estimate of noise from noisy signal [27]. The noise power spectrum estimation is obtained by averaging over multiple frames of a known noise segment; which can be detected using voice-activity detector (VAD) [28]. Assuming an additive model of noise, and given the linearity property of the Fourier transform of the noisy signal, clean signal, and noise (respectively), [29] we get:

$$Y(e^{wj}) = X(e^{wj}) + N(e^{wj}) \quad (1)$$

4. Approaches for Speech Enhancement

In a novel approach, HMM-based and USS synthesized speech is used for speech enhancement. This means that the initially spoken sentence is re-synthesized to enhance the audio. To increase the quality of the generated signal, spectral subtraction is used. The following section defines the HMM and USS approach for speech enhancement.

4.1. Hidden Markov Model

In HMM approach, the combination of Log F0 and Mel-generalized cepstrum coefficients (MGCC) coefficients of the

synthesized speech signal, from the MLSA filter is noted down [30]. The variation is found in these two parameters from the original speech. This means that the excitation signal for one phoneme, but the coefficients for a different one are used, resulting in faulty speech. To avoid this mismatch, forced alignment of the two components is applied. For this purpose, labels are given to the model which then forcefully tries to find their temporal appearance in the audio and adds them to the label, similar to timestamps. Thus, coefficients can be generated which match the course of the pitch frequency [31]. If interferences are present in the audio, however, the forced alignment is negatively affected as well. Thus, the original speech signal is re-synthesized and enhanced. In this the gamma tone filtering is applied to the synthesized TTS voice and then using spectral subtraction technique the voice is re-synthesized [32]. Figure 2 explains the speech enhanced approach for HMM-based Marathi TTS system.

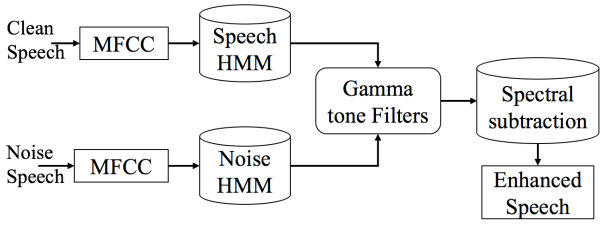

Figure 2: *Block diagram of the proposed speech enhancement for HMM*

4.2. Unit Selection Synthesis

The Marathi speech is synthesized using USS from the unit database built using Marathi recordings. Units (in the current work, phonemes) are selected to produce a natural realisation of a target phoneme sequence predicted from the text which is annotated with prosodic and phonetic context information [13]. The system is trained to build the TTS with related to speech sound units vowel like regions (VLR) and non-vowel like regions (NVLR) that are present in speech corpus. Vowel, diphthong, semivowel and nasal sound units are referred to VLR. Stop consonants, fricatives and affricates are classified as NVLR. The target speech produced using VLR is smoothly generated using USS method [33]. The glitches are observed during the synthesizing of consonant clusters. Thus to improve the quality of synthesized voice the gamma tone filtering method is applied. Then using spectral subtraction technique, the more enhanced TTS speech is produced [32]. The block diagram of the proposed speech enhancement method for USS Marathi TTS system is shown in Figure 3.

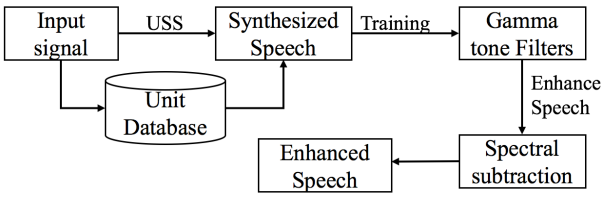

Figure 3: *Block diagram of the proposed speech enhancement for USS*

4.3. Marathi Database

In this section, we discussed the various steps in building the speech corpus. The recording media is chosen so as to capture the effects of the microphone. The sentences for recording are chosen from regular novel books. We selected from the Marathi corpus 1 speaker with same accent region (Maharashtra, India) and another is the TTS Speech. There are around 1000 sentences available from the speaker which comes around 3 hours of recorded speech. All data is sampled at 48 kHz and orthographic transcription is also available [34].

5. Experiments and Results

Initially, the system is trained with USS and HMM-based techniques. The output voice is re-synthesized to improve the quality using spectral subtraction method. The spectral properties of the signal are studied after re-synthesizing the output speech. The figure below shows the difference in original, synthesized and enhanced TTS speech respectively for both USS and HMM techniques [32]. The figure below contains the spectral properties for sentence "***kaahii vishishhat:a dhayeiyaanei pareiriita aajakaala phaaracha kamii lookan: saapad:atii***l". The graph shows that with the enhanced TTS, the production of phonemes is much more improved as compared to other two methods.

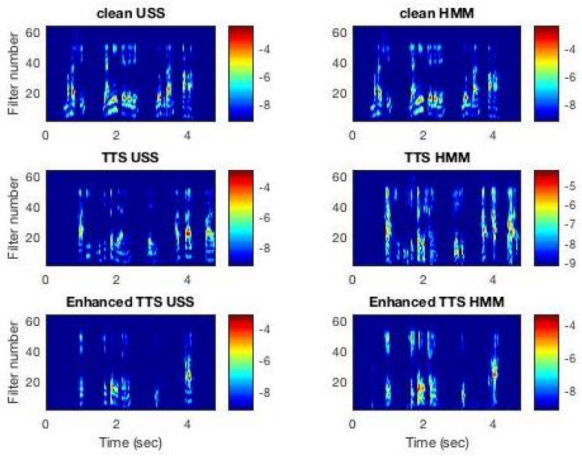

Figure 4: *Spectrum modelled of noise speech in HMM and USS by natural, synthesis and enhanced*

6. Conclusions

The synthesized speech is evaluated on basis of two parameters i.e. naturalness and intelligibility. The paper focused on re-synthesize TTS speech using spectral subtraction for the system that is initially trained with both USS and HMM techniques for low resource Marathi language. Speech enhancement technique provides a better quality of synthesized speech. The study of spectral properties of original and enhanced TTS shows very promising results, with a clear increase of quality and intelligibility as compared to normally generated speech.

7. Acknowledgements

We are thankful to Dr. Bharti Gawali Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing the guidance for the research.

8. References

[1] W. Holmes, *Speech synthesis and recognition*. CRC press, 2001.

[2] P. Marler, "Birdsong and speech development: Could there be parallels there may be basic rules governing vocal learning to which many species conform, including man," *American scientist*, vol. 58, no. 6, pp. 669–673, 1970.

[3] A. Balyan, S. Agrawal, A. Dev, and R. Kumari, "Development and implementation of hindi tts," in *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*. IEEE, 2015, pp. 458–463.

[4] B. Langner and A. W. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Pennsylvania, USA.*, vol. 1, 2005, pp. I–265.

[5] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMM with dynamic features," in *EEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Atlanta, GA, USA*.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Washington, DC, USA*, vol. 4, 1979, pp. 208–211.

[7] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing, Elsevier*, vol. 41, no. 1-4, pp. 1–24, 2001.

[8] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for blizzard challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop, Pittsburgh PA*, 2006.

[9] E. Ambikairajah, J. Epps, and L. Lin, "Wideband speech and audio coding using gammatone filter banks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, UT, USA*, vol. 2, 2001, pp. 773–776.

[10] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Detroit, MI, USA*, vol. 1, 1995, pp. 660–663.

[11] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the wiener filter for noise reduction," in *Speech Enhancement, Springer*, 2005, pp. 9–41.

[12] T. Dutoit, *An introduction to text-to-speech synthesis*. Springer Science and Business Media, 1997, vol. 3.

[13] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Atlanta, GA, USA*, vol. 1, 1996, pp. 373–376.

[14] A. W. Black and P. A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," pp. 1–4, 1997.

[15] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis." pp. 1–4, 1995.

[16] N. Campbell and A. W. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in speech synthesis, Springer*, 1997, pp. 279–292.

[17] R. Ribeiro and I. Oliveira, Trancoso, "Using morphossyntactic information in tts systems: Comparing strategies for european portuguese," in *International Workshop on Computational Processing of the Portuguese Language, Springer*, 2003, pp. 143–150.

[18] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication, Elsevier*, vol. 9, no. 5-6, pp. 453–467, 1990.

[19] N. S. Krishna, H. A. Murthy, and T. A. Gonsalves, "Text-to-speech (TTS) in indian languages," in 1^{st} *International Conference on Natural Language Processing at NCST, CDAC-Mumbai)*, 2002, pp. 317–326.

[20] M. B. Mustafa, Z. M. Don, and G. Knowles, "Context-dependent labels for an hmm-based speech synthesis system for malay hmm-based speech synthesis system for malay," in *International Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[21] S. J. Lee, B. O. Kang, H. Chung, and J. G. Park, "A useful feature-engineering approach for a lvcsr system based on cd-dnn-hmm algorithm," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1421–1425.

[22] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[23] A. Raux and A. W. Black, "A Unit selection approach to F0 modeling and its application to emphasis," in *IEEE Workshop on Automatic Speech Recognition and Understanding, St Thomas, VI, USA*, 2003, pp. 700–705.

[24] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop, Santa Monica, CA, USA*, 2002, pp. 227–230.

[25] S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration Hidden Markov Model," *IEEE signal processing letters*, vol. 10, no. 1, pp. 11–14, 2003.

[26] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in 6^{th} *European Conference on Speech Communication and Technology, Lisboa, Portugal*, 1999, pp. 1–4.

[27] E. Verteletskaya and B. Simak, "Noise reduction based on modified spectral subtraction method," *IAENG International journal of computer science*, vol. 38, no. 1, pp. 82–88, 2011.

[28] H. Veisi and H. Sameti, "Hidden Markov Model-based voice activity detector with high speech detection rate for speech enhancement," *IET signal processing*, vol. 6, no. 1, pp. 54–63, 2012.

[29] Takayoshi, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in 5^{th} *International Conference on Spoken Language Processing, Sydney, Australia*, 1998, pp. 1–4.

[30] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in 3^{rd} *International Conference on Spoken Language Processing, Berlin, Germany*, 1994, pp. 1–4.

[31] A. Kato, "Hidden Markov Model-based speech enhancement," Ph.D. dissertation, University of East Anglia, 2017.

[32] S. Kayte, "A Text to Speech System for Marathi using English language," *International Journal of Engineering Science and Generic Research*, vol. 1, no. 1, pp. 1–12, 2015.

[33] S. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 8, pp. 2552–2565, 2011.

[34] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H INDIC speech databases," in 13^{th} *Annual Conference of the International Speech Communication Association, Portland, Orego*, 2012, pp. 1–4.