International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-6 Issue-5, June 2017

Speaker Independent Text to Speech for Malayalam

Sajini T, Neetha George

Abstract—Text to speech (TTS) relates is software which converts text to speech output. TTS has wide range of applications which includes assistive technologies like communication devices for providing voice for voice disabled. These applications require flexibility to provide diverse speakers voice or unique voice as output. Existing corpus based TTS does not provide this flexibility, and changing a voice is time consuming, expensive and tedious since it requires hours of high quality speech corpus. In this work we explore the speaker adaptation technology available in Hidden Markov Model based Text to speech (HTS) for providing speaker variability in Malayalam TTS. Speaker adaptation (SA) using HTS framework has been successfully implemented for foreign languages like English, Japanese etc. but not yet been tried for Indian languages. In this work we try to implement SA using HTS framework as a solution for providing diverse voices, reducing the expenses, time and effort required, in the usual approach for creating a variant/new TTS voice. We have used a combination of the constrained maximum likelihood linear regression (CMLLR) and maximum a posterior probability (MAP) for generating variant voices. A five speaker database with one hour speech from each speaker is used for SA, in which four speakers database is used for training speaker independent average model (SI). SI model was trained with different number of speakers. Average model with 3 speakers gave an intelligible noisy output, and four speakers gave intelligible, good quality and similarity output with rarely occurring distortions. Quality of the system was determined using perceptual scores tested with 15 native speakers. An average word error rate (WER) for 3 and 4 speaker model was 15.65% and 16.2% for paragraphs selected from different domains and 30 sentences gave an average score of 26.82% and 21.14%. The adapted voice model gave a 3.39, 3.59, 3.55 and 3.38 as the Mean opinion score (MOS) for naturalness, intelligibility, degradation and similarity index. The results show that the SA technique for HTS is a quick, easy & less expensive technique that can be successfully used for a phonetic language like Malayalam for providing generating diverse voices for TTS.

Index Terms: Speaker adaptation, HMM based TTS, Constrained maximum likelihood linear regression, Maximum a posterior, MAP.

I. INTRODUCTION

Text to speech (TTS) is software which converts text input into synthetic speech. Good quality TTS, in wide choice of voice is available for foreign languages, but very few is available in the Indian Languages. For Malayalam their exist TTS in the state of the art technologies like corpus based Unit selection synthesis(USS) and Statistical parametric synthesis based on HMM based text to speech (HTS). Even though TTS exist for Malayalam choice of voices are limited due to the expensive and time consuming process of developing database for a new language. In many TTS applications, it is required to have choice of voices, and in some cases personalization of the TTS voice is preferred by the user. Choice of voices can be made available only if there is availability of TTS in diverse voices. Voice conversion or voice adaptation is the available can address these limitations in TTS.

Voice conversion/ Speaker adaptation is a technique that modify a source voice to the target speaker voice, which also helps in personalizing speech synthesizer. The source and the target differ in terms of voice characteristics. This feature is highly required for applications using TTS, especially in developing Augmentative alternative communication (AAC) applications. In case of Amyotrophic Lateral Sclerosis (ALS), people lose their ability to speak after acquiring language knowledge. Such people with progressive speech loss can use their voice data base to create their own voice in AAC.

There are different approaches for voice adaption, which includes signal based and parametric based.

In Voice track length normalization is normally done in speech recognition to separate the speaker individual characteristics. Frequency warping technique [11] is has been used in order to reduce the weighted spectral distance between the source speaker and the target speaker. Voice conversion can be done in two ways either by conversion of the parameters to the target parameter or by modifying the output i.e. speech to the target requirements. The other techniques include Vector quantization mapping cook books, dynamic time wrapping, neural network and Gaussian mixture model. Statistical parametric approaches to speech synthesis (such as hidden Markov model (HMM)-based speech synthesis) have grown in popularity over the last few years [12]. These approaches have been shown to be less sensitive than unit selection to imperfect training data [13]. In HMM-based speech synthesis, use of parameter sharing techniques allows the synthesis of models for speech units unseen in the training corpus; this contrasts with the corresponding strategy that must be used in unit selection where the system must select a substitute unit, typically on the basis of heuristics [5]. SPS using HMM provides flexibility to adapt, already-trained HMM-based systems to the voice characteristics of a target speaker with small amounts of adaptation data.

Manuscript published on 30 June 2017.
***** Correspondence Author (s)
Sajini T, Electronics and communication, Kerala university, College of Engineering, Trivandrum, India, E-mail:sajnathattankandy@gmail.com
Neetha George, Electronics and communication, Kerala university, College of Engineering, Trivandrum, India, E-mail: neethabj@gmail.com

The fact that it is possible to use HMMs that have been trained on cleanly recorded data, rich in phonetic contexts, as the basis for adaptation means that high-quality speech. can be synthesized even when the adaptation data is noisy and sparse[5]. Voice adaptation can be used to change the characteristics like speaking style, dialect etc. into the existing SI model. Even though many literatures on text dependent, text independent approaches are available for foreign languages, non-have implemented or tried for Indian Languages. This work we focus on developing an HTS voice model for the existing Malayalam TTS, using the speaker adaptation technique available in the HTS framework.

The HTS voice adapt package is customized and used for voice adaptation in this work. We have used constrained maximum likelihood linear regression (CMLLR) and maximum a posterior (MAP) for adapting voice models which include spectral, excitation and duration models for Malayalam. A corpus of 5 speakers was selected for this work, in 4 speakers data was used for training different speaker independent models and one speaker was identified as the target, to which the average speaker independent models are adapted. The database selected contains 5K sentences with more than 90% of the total phone coverage for Malayalam. The section 2 covers literature review, section 3 covers the details of database preparation for Malayalam, Section 4 covers the voice adaptation for Malayalam using HTS based system, section 5 covers the results, and section 6 covers the conclusion and future scopes.

In this work we have used a text dependent approach for voice adaptation. But existing system has a limitation of speaker dependence and unable to provide voice in user choice. It is difficult to change the voice characteristics in the existing system. A text dependent approach in which we uses the speech corpus created with the same text corpus has been used for voice adaptation. The adapted voice model is then integrated with exiting HTS based TTS and quality assessment is done using perceptual scoring.

Even though speaker adaption has been done for other languages; similar work has not been done for Malayalam. In the current work, voice adaptation done using CMLLR and MAP has achieved a WER, DMOS, MOS and SI s a 3.39, 3.59, 3.55 and 3.38 respectively. This work will provide a practical solution for implementing flexibility of providing diverse voice and personalization in SPS based TTS.

II. LITERATURE REVIEW

The development of computer-based speech synthesis technology has been ongoing for decades. In the early days, rule-based synthesis dominated the speech synthesis research. It generates synthetic speech by manipulating speech segments according to handcrafted rules. In 1990s, the speech synthesis technology progressed from the rule-based approach to the data-driven, corpus based one. High-quality speech synthesizers can be built from sufficiently diverse single-speaker speech databases. We can see progress from fixed inventories, found in diphone synthesis, to the more general techniques of unit-selection synthesis, where appropriate sub word units are selected from large databases. Unit-selection techniques evolved to be the dominant approach to speech synthesis.

The idea of HMM-based speech synthesis first appeared in mid-1990s. It has been popular in speech synthesis research since the early 2000s [28] [53], the basic procedure of HMM-based speech synthesis is described in this work. Statistical parametric synthesis (such as hidden Markov model (HMM)-based speech synthesis) has grown in popularity over the last few years [1]. These approaches have been shown to be less sensitive than unit selection to imperfect training data [2]. Unlike unit selection synthesis, already-trained HMM-based systems can be adapted to the voice characteristics of a target speaker with small amounts of adaptation data [5]. Adaptation has been used to impose various types of characteristics onto existing statistical parametric synthesizers, for example, characteristics associated with dialect [3] and speaking style [4]. The different features of HMM based TTS (HTS) include transforming voice characteristics, speaking styles and emotions. The main advantage of statistical parametric synthesis (SPS) is its flexibility in changing its voice characteristics, speaking styles, and emotions [1] [9].

The voice conversion in the area of speech processing, speech parameterization and modification, standalone voice conversion, basic approaches, problems and improvements in GMM based conversion like over fitting over smoothing, time dependent mapping, and advanced code book based methods and SPS based voice conversions are well documented [52]. Voice transformation using Source modifications, filter modifications, combining source and filter modification's is detailed in [49]. MFCC feature extractions for parameterization of speech and speech analysis for digitizing and to produce voice features are detailed in [44]. Mel generalized cepstral coefficient approach (MGC) is a generalized cepstral analysis method which is used in HTS. MGC is viewed as a unified approach to the cepstral method and the linear prediction method [40]. The database for speaker adaptation for Malayalam is prepared based on CMU arctic database [47], and using language specific rules for Malayalam [45] [27] [46]. In HTS based system the spectrum pitch and duration are simultaneously modeled [48].

Two major techniques in adaptation are maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994) and maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995). MAP estimation involves the use of prior knowledge about the distributions of model parameters. Hence, if we know what the parameters of the model are likely to be (before observing any adaptation data) using prior knowledge, we might well be able to make good use of the limited amount of adaptation data. A major drawback of MAP estimation is that every Gaussian distribution is individually updated. If the adaptation data are sparse, then many of the model parameters will not be updated. This causes the speaker characteristics of synthesized speech to often switch between general and target speakers within an utterance.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Various attempts have been made to overcome this, such as vector field smoothing (VFS) (Takahashi and Sagayama, 1995) and structured MAP estimation (Shinoda and Lee, 2001). Adaptation can also be accomplished by using maximum likelihood linear regression (MLLR).

In MLLR, a set of linear transforms is used to map an existing model set into a new adapted model set such that the likelihood for adaptation data is maximized. The state-output distributions are usually clustered by a regression-class tree, and transformation matrices and bias vectors are shared among state-output distributions clustered into the same regression class (Gales, 1996). By changing the size of the regression-class tree according to the amount of adaptation data, we can control the complexity and generalization abilities of adaptation. There are two main variants of MLLR. If the same transforms are trained for A and H, this is called constrained MLLR (or feature-space MLLR); otherwise, it is called unconstrained MLLR (Gales, 1998). For cases where adaptation data are limited, MLLR is currently a more effective form of adaptation than MAP estimation. Furthermore, MLLR offers adaptive training (Anastasakos et al., 1996; Gales, 1998), which can be used to estimate "canonical" models for training general models. For each training speaker, a set of MLLR transforms is estimated, and then the canonical model is estimated given all these speaker transforms. Yamagishi applied this MLLR-based adaptive training and adaptation techniques to HMM-based speech synthesis (Yamagishi, 2006). This approach is called average voice-based speech synthesis (AVSS). It could be used to synthesize high-quality speech with the speaker's voice characteristics by only using a few minutes of the target speaker's speech data (Yamagishi et al., 2008b). Furthermore, even if hours of the target speaker's speech data were used, AVSS could still synthesize speech that had equal or better quality than speaker-dependent systems (Yamagishi et al., 2008c). Estimating linear-transformation matrices based on the MAP criterion (Yamagishi et al., 2009) and combining MAP estimation and MLLR have also been proposed (Ogata et al., 2006). The use of the adaptation technique to create new voices makes statistical parametric speech synthesis more attractive. Usually, supervised adaptation is undertaken in speech synthesis, i.e., correct context-dependent labels that are transcribed manually or annotated automatically from texts and audio files are used for adaptation. Voice adaptation can also be done using MLLR techniques **[10]**. An arbitrary speaker Characteristics using speaker independent speech units, "average voice" units [51]. HTS models state duration, [18], state durations modeled by a multi-dimensional Gaussian distribution, and duration models are clustered using a decision tree based context clustering technique. For handling the unseen units HTS uses a method of creating a tied-sate using phonetic decision tree [27]. In order to simultaneously model and adapt excitation parameters of speech as well as spectral parameters, the multi-space probability distribution (MSD) HMM and its MLLR adaptation algorithms are used. The logarithmic fundamental frequency (log F0) is used as the excitation parameter. The MSD-HMM enables to treat the log F0 observation, which is a mixture observation of a one-dimensional real number for voiced regions and a symbol string for unvoiced regions, within a generative model. In order to simultaneously model and adapt duration parameters for the spectral and excitation parameters as well, MSD hidden semi-Markov model (MSD-HSMM) and its MLLR adaptation algorithm are used. The HSMM is an extended HMM, having explicit state duration distributions instead of the transition probabilities to directly model and control state durations. More advanced speaker adaptation techniques including constrained structural maximum a posteriori linear regression (CSMAPLR) For evaluation of the system the five point perception score (MOS) [9], [24] is used as the evaluation scale. Perceptual intelligibility (PI), similarity scores and their MOS rating scales are given in these works.

III. DATABASE

One of the prime requirements of speech technology development is a large database, which represents the language. The database must cover all phones, allophone and must be phonetically balanced. For developing TTS the quality of database is very important factor. For voice adaptation an average SI model is to be created first. This requires multiple speaker databases, covering the wide variation in the language. Database can be created with the same text corpus (text dependent) or with different text corpus (text independent). In this work we have used text dependent corpus of experimenting voice adaptation. HTS voice adaptation using arctic database is used as the reference and customized for modeling SI TTS for Malayalam [21]. In this sample they have used 6 speakers, text dependent database for generating the SI model. To build a HTS based system for a new language requires phone set, question set, Letter to sound (LTS) rules. The inputs required for training SI model are speech corpus(wav format, text dependent, multi speaker) and the transcription, that are the time aligned phonetic transcription is generated using festival framework[26]. The computer pores through the database, i.e. the audio files and their transcriptions and models the speech. Database preparation is costly, time consuming work. For this work text dependent corpus is used, which is created from the corpus available at CDAC-T. One hour database of five speakers (one female and four male) database is collected for creating the multi speaker database for training SI model. Waves and transcription are available for this database. Pronunciation dictionary is prepared by selecting word, from the available pronunciation lexicon using Perl parser.

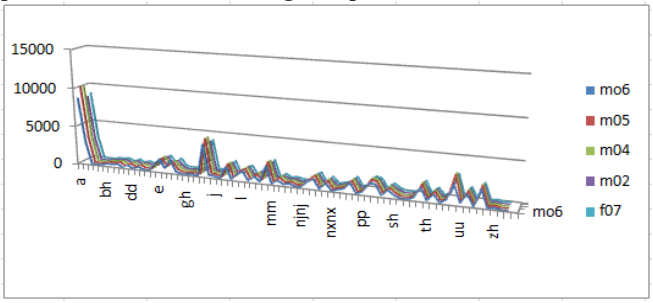

Fig 3.1: Phone variations in database

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

BBBThe recording specifications are PCM, sampling rate 48 KHz. The speech data is collected from 5 speakers, one female and four male speakers in an acoustically treated room.

IV. VOICE ADAPTATION FOR MALAYALAM USING HTS

For developing the speaker independent TTS, the first step is the training for developing average voice model. The details of training the average model is given below.

A. Setting up the working environment for voice adaptation

Download and install the voice adaptation packages and its dependencies.

B. Generate the data for training using the speech database

First step is converting the speech database in wave format to raw. The time aligned utterances is generated using festival framework. Along with wave and the utterances, phone features are also given as input for HTS for training.

C. Parameter modification in HTS frame work for the Training

The parameters like pitch, frequency wraping factor is modified for selected speaker. The pitch range for each speaker is identified using wavesurfer. The minimum and maximum values of pitch is calculated for the sample selected from each speaker. Generate mgc (Mel generalized cepstral coefficients), lf0 (log fundamental frequency) and context dependent labels from time aligned transcriptions. These are the input for training the average model, which will be adapted to the target voice.

D. Training for creating SI model for speaker adaptation

Compute the global variance (GV). GV is computed from the original wav is used in synthetic voice generation for removing over smoothening in output. HComp command in HTK is used for computing variance floor.

E. Create the initial models

Context independent models for each phone are trained. Initial HMM CI models are generated from the mgc and lf0. Each phone is modeled as a multi stream HMM models, with 5 states each with 4 streams and simultaneously models spectral and excitation. First stream stores mgc (105 dimensional vectors). The mean and variance, their delta and double delta coefficients are stored. Spectral information is stored as a multivariate Gaussian distribution. Excitation is a 3 dimensional vector, for log fundamental frequency and weight for voiced and unvoiced is stored. HInit is used for creating initial models and HRest for re estimating the parameters for maximizing the likelihood of initial models. HHEd is used to form a macro file containing all the monophone models.

F. Parametric re-estimation

The parameters are re-estimated using EM algorithm. This process maximizes the probability for getting the given observation. HERest is used for embedded re-estimation.

G. Create context dependent models

Context dependent state models are generated using CD labels. It copies the CI models and concatenate, using CD labels to CDHMM. Training of CDHMM is done using the Observation sequence. Parameters are then re-estimated using EM algorithm for CDHMM models.

H. Decision tree based context clustering

Decision-tree-based context clustering was applied to spectral, f0 (aperiodicity) and duration features separately. Clustered parameters were tied and re-estimated, untie and re-estimated. This clustering, tying, re-estimating are done depending on the iterations set.

- Stream dependent tree based clustering
- Spectrum and excitation have different context dependencies
- Build decision tree separately
- Re estimation of CDHMM using EM
- Estimated HMM models
- HHEd and HERest are used for clustering mgc, lf0 and duration models.

I. Duration modeling

Duration modeling is done as a single pdf distribution

- Estimate context dependent duration models
- Decision tree based clustering is done
- Estimate duration models

J. Parameter tying

State tying is done to cluster similar states and to tie model parameters among several context-dependent HMMs so that we can estimate model parameters more robustly. The state tying process is conducted in a hierarchical tree structure manner, and the tree size is automatically determined based on an information criterion called minimum description length (MDL). As the spectral, excitation, and duration parameters have different context dependency, they are clustered separately by using stream-dependent decision trees

- Untying and re estimation of parameters
- Untie the parameter sharing structure using HHEd and perform embedded re-estimation using HERest.
- Re clustering and re-estimation

K. Average voice model

The average model generated by the training module will be transformed to get a speaker independent model. Adaptation of speaker independent model has the following advantages

- Average model gives better accuracy than single speaker model
- Reduces bias
- SI training gives the best set of model parameters

L. HTS speaker adaptation

In HTS voice adapt, speaker adaptation is done using a combination of CMLLR and MAP

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

- MLLR performs linear transforms of mean vectors of the state output probability distributions. The covariance matrices as well as mean vectors of the state output probability distributions are transformed using the same matrices, this is called constrained MLLR
- These transforms may be estimated using the standard maximum likelihood or MAP criteria.
- (adaptation of HMM/GMM parameter), and may be combined with other speaker adaptation techniques such as vocal tract length normalization

For creating a HTS voice model for Malayalam TTS, 5 speaker databases are used. Four speakers were selected for training SI models, and one speaker is identified as target. Training has been done for creating 1, 2, 3, 4 speaker SI models. SAT for four speaker model took 240 hours for completing the training and adaptation. Pitch value was set as in the table 3.1. & frequency warping factor alpha is set as 0.55 for all the training.

V. RESULTS

This work explored the possibility of the voice conversion/adaptation feature in HTS based system, for Malayalam. The quality of TTS is analyzed based on perception. To evaluate the quality of the HTS model created for speaker independent TTS, listening tests are conducted. The following perceptual tests were conducted to evaluate the system, mean opinion score (MOS) for intelligibility and naturalness, Degradation MOS (DMOS), Similarity test and Word error rate (WER).

The MOS is the simplest method to evaluate the quality of a speech synthesis system. The five point scale for MOS is 5-Excellent, 4- Good, 3-Fair, 2-Poor and 1-Bad.

The five point MOS scale for evaluating naturalness is 5-System sounds like human, 4-Robotic sound but reading correctly, 3-Reading sentences with less broken words in robotic manner, 2-Almost every word broken, 1- Extremely intolerable.

The average MOS is calculated as in equation 2

$$\text{MOS total sentences} = \frac{\sum_{j=1}^{N} \frac{\sum_{n=1}^{N} Rn}{N}}{M} \quad \ldots\ldots (2)$$

Sentence index j, and M is the # of sentences

The five point scale for DMOS is defined as 5-In audible, 4-Audible but annoying, 3-Slightly annoying, 2-Annoying and 1-Very annoying.

The Similarity is measured based on the five point perception scale as 5-Both speakers are clearly the same, 4-Both speakers are probably the same, 3-Cannot determine whether both speakers are the same or not, 2-Both speakers are probably different and 1-Speakers are clearly different.

To evaluate the quality of voice adaptation for generating variant voices were done using the perceptual evaluation technique detailed in the previous chapter. Training the average model and voice adaptation was done using one, two, three and four speaker database. This was done to identify the minimum requirement for average model to generate intelligible less distortion voice output. Details of the experiments result are given below.

A. Minimum speaker requirement of average model – speaker dependent (SD) model

Among the five speakers in the database, different models were generated using the HTS framework. Average model were trained using different number of speakers. Initially a single speaker voice is trained and adapted to target speaker. The output quality was not acceptable. The best result was found for an average model with 4 speakers. The details of observations for the different models created with different number of speaker data are as given in the table 5.1.

Table 5.1: Details of models trained for speaker adaptation

Model	Quality		Remarks
Voice	**Intelligibility**	**Discontinuity**	
Single	Poor	Frequent	SD high, so voice create more distortions
Two	Poor	Frequent	Model mapping inaccurate
Three (Single gender)	Intelligible	Noisy	Model accurate context dependent
Four (Mixed gender)	Intelligible, good quality, good similarity	Rarely	Average model give output with high intelligibility and similarity score

B. Perception score

For evaluation and conducting listening test, 30 sentences and 20 paragraphs, which were selected from different domains, were used. Native speakers were selected as evaluators for the listening test. The audios were played and asked the evaluators to score the sentences, based on the above scales. To test the model accuracy, sentences were synthesized using the two models (3rd and 4th) listed in the above table was used.

C. Word Error Rate (WER)

WER is evaluated for the two HTS models single gender and mixed gender, using the paragraphs and sentences. The WER for sentences synthesized for 20 paragraphs selected from different domain, using the two models are given in table 6.6. The WER for 30 sentences covering the four domains is given in table 5.2.

D. Perception score for Sentences

Perception score for sentences were done based on the listening tests carried out with the native speakers. The native speakers were made to listen to the 20 sentences and rate the sentences based on the MOS scales defined for intelligibility, naturalness, similarity and DMOS.

The final score is calculated using equation 2, which gives the average MOS score for the tests.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 5.2: WER for different Paragraphs selected from different domain

Category	NEWS	STORY	SPORTS	GENERAL
Singe Gender Average WER %	18.15	14.22	12.69	17.53
Average WER%	15.65			
Mixed Gender Average WER %	14.05	16.21	14.97	19.23
Average WER%	16.12			

Table 5.3: WER rate for sentences

	Sentences	Typo errors removed
Single gender	32.27	26.82
Mixed gender	22.03	21.14

Table 5.4: gives the average score for MOS, DMOS and SI

	MOS-I	MOS-N	DMOS	SI
Single gender	3.52	3.39	3.38	3.37
Mixed gender	3.68	3.59	3.55	3.42

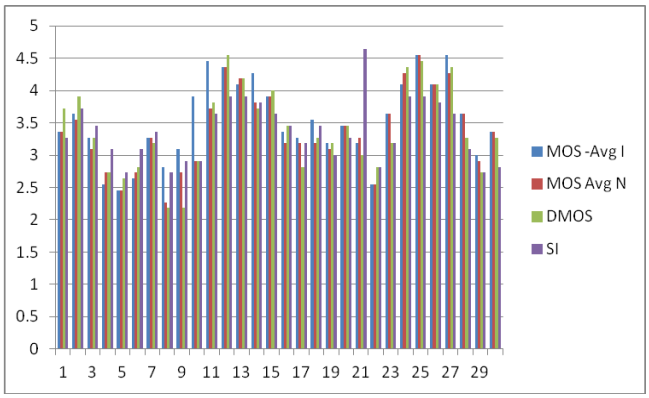

Figure 5.1: Average MOS-I, MOS-N, DMOS and Similarity score for single gender – 3 speakers

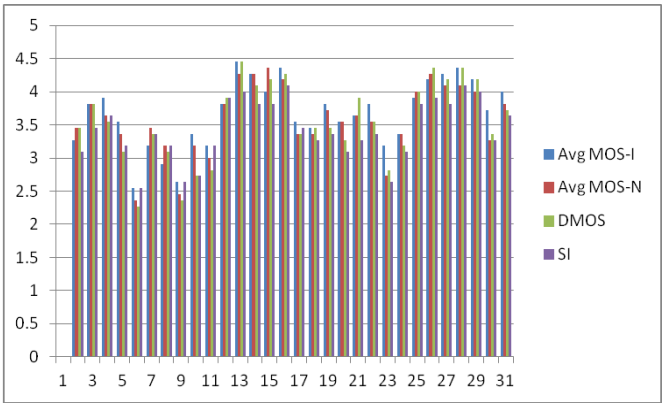

Figure 5.2: Average MOS-I, MOS-N, DMOS and Similarity score for single gender – 4 speakers

Table 5.8: gives the average score for MOS, DMOS and SI

	MOS-I	MOS-N	DMOS	SI
Single gender	**3.52**	**3.39**	**3.38**	**3.37**
Mixed gender	**3.68**	**3.59**	**3.55**	**3.42**

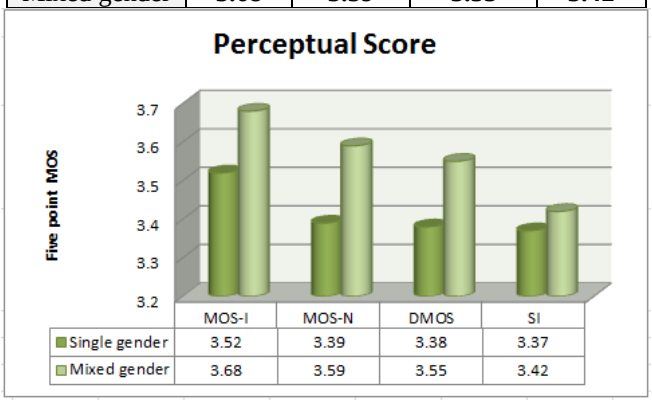

Figure 5.3: Average perceptual score for MOS-I &N, DMOS and SI

For evaluating the quality of the HTS voice model, 30 sentences were selected and synthesized. Native speakers were identified for evaluating the sentences. They were made to listen to the sentences played and score the sentences on the 5 point scale for MOS, DMOS and SI. Fig 6.1 & 6.2 shows the average rating for MOS, intelligibility and naturalness, DMOS and similarity score for the sentences using both single gender and mixed gender. The table 6.8 gives the average rating of the above scores.

The results show that the quality of the adapted voice improves as the number speaker voice used for creating average model increases. Over all an average of 10% improvement in observed in the speaker adaptation done with 4 speakers.

VI. CONCLUSION AND FUTURE SCOPE

In this work we implemented Speaker adaptation using HTS framework for Malayalam which provides a successful solution for reducing speaker dependence in TTS. The perceptual results shows that it can be used as a solution for providing diverse voices, reducing the expenses, time and effort required, in the usual approach for creating a variant/new TTS voice. We have used a combination of the constrained maximum likelihood linear regression (CMLLR) and maximum a posterior probability (MAP). A five speaker database with one hour speech from each speaker is used for SA, in which four speakers database is used for training speaker independent average model (SI). SI model was trained with different number of speakers. Average model with 3 speakers gave an intelligible, noisy output. With four speakers the models were more accurate and gave an output with better context dependence. The output was observed to be with good quality and similarity, with rarely occurring distortions. The perceptual scores were used to evaluate the quality of the adapted voice model. The system was evaluated with 15 native speakers.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

An average WER of 15.65% was observed for 3 speaker models and 15.65% was observed for 4 speaker model for the selected paragraphs. For sentences the average scores 26.82% and 21.14%. The result shows that the quality of adaptation has correlated with the numbers of speakers covered for creating the SI model.

The MOS score for Intelligibility, naturalness, degradation and similarity score for the two models with 3 speakers was observed as 3.52, 3.39, 3.38 and 3.37. The above score for 4 speakers was observed to be 3.68, 3.59, 3.55 and 3.42. The MOS shows that the quality improves as the number of speakers used for creating average model increases, since the speaker dependence is reduced and a more generalized acoustic models are created.

Database phone frequency also affects synthesis quality. So for average model creation database must be prepared taking into account of the phone coverage, especially for the high frequent vowels. In this work we have used database with phone coverage of 72% percentage. Even though missing phones were intelligible, quality is affected. Quality of phones depends on the phone coverage in the database, and depends on the frequency of occurrence covering different contexts.

As an extension to this work, we plan to build a corpus ensuring the phone coverage, in different context with multiple speaker and multiple dialects. SA quality can be improved with this database since the model dependence can be reduced. We also plan to try text independent approach in which speech corpus created with different text corpus can be used for training the SI model. We also plan to implement unsupervised training, with voice recognition and also to explore rapid speaker adaptation using Eigen vectors.

ACKNOWLEDGMENT

I would like to thank Mr. V K Bhadran, Associate director CDAC and Dr. Deepa P Gopinath for their valuable suggestion and guidance.

REFERENCES

1. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039-1064.
2. J. Yamagishi, J., Ling, Z., & King, S. (2008). Robustness of HMM-based speech synthesis.
3. J. Yamagishi, J., Zen, H., Wu, Y. J., Toda, T., & Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge.
4. J. Yamagishi, J., Tachibana, M., Masuko, T., & Kobayashi, T. (2004, May). Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on (Vol. 1, pp. I-5). IEEE.
5. Watts, O., Yamagishi, J., King, S., & Berkling, K. (2010). Synthesis of child speech with HMM adaptation and voice conversion. IEEE Transactions on Audio, Speech, and Language Processing, 18(5), 1005-1016.
6. Yamagishi, J., Zen, H., Toda, T., & Tokuda, K. (2007). Speaker-Independent HMM-based Speech Synthesis System: HTS-2007 System for the Blizzard Challenge 2007.
7. Yamagishi, J., & Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE TRANSACTIONS on Information and Systems, 90(2), 533-543.
8. J Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE TRANSACTIONS on Information and Systems, 88(3), 502-509.
9. Latorre, J., Iwano, K. & Furui, S. (2006). New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. Speech Communication, 48, 1227-1242.
10. Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (1998). Speaker adaptation for HMM-based speech synthesis system using MLLR. In the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.
11. Black, A. W., Zen, H., & Tokuda, K. (2007, April). Statistical parametric speech synthesis. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 4, pp. IV-1229). IEEE.
12. J Yamagishi, J., Ling, Z., & King, S. (2008). Robustness of HMM-based speech synthesis.
13. Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A., & Narayanan, S. (2006, May). Text-independent voice conversion based on unit selection. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Vol. 1, pp. I-I). IEEE.
14. MOULINES, E. (1995). Voice Conversion: State of Art and Perspectives. Speech Communication, 16, 125-126.
15. https://www.iitm.ac.in/donlab/tts/cls.php "Indian Language Speech sound Label set (ILSL12) Version 2.1.6"
16. Fukada, T., Tokuda, K., Kobayashi, T., & Imai, S. (1992, March). An adaptive algorithm for Mel-cepstral analysis of speech. In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on (Vol. 1, pp. 137-140). IEEE.
17. Tokuda, K., Masuko, T., Miyazaki, N., & Kobayashi, T. (2002). Multi-space probability distribution HMM. IEICE TRANSACTIONS on Information and Systems, 85(3), 455-464.
18. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1998, December). Duration modeling for HMM-based speech synthesis. In ICSLP (Vol. 98, pp. 29-32).
19. K Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (Vol. 3, pp. 1315-1318). IEEE
20. Yamagishi, J., Zen, H., Toda, T., & Tokuda, K. (2007). Speaker-Independent HMM-based Speech Synthesis System.
21. Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (1998). Speaker adaptation for HMM-based speech synthesis system using MLLR. In the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.
22. http://www.ling.upenn.edu/courses/Spring_2001/ling001/phonology.html
23. http://www.phon.ucl.ac.uk/courses/spsci/b214/week2-5.pdf
24. http://tdil-dc.in/undertaking/article/449854TTS_Testing_Strategy_ver_2.1.pdf
25. Ramani, B., Christina, S. L., Rachel, G. A., Solomi, V. S., Nandwana, M. K., Prakash, A., ... & Vijayalakshmi, P. (2013). A common attribute based unified HTS framework for speech synthesis in Indian languages. In Eighth ISCA Workshop on Speech Synthesis.
26. Kumar, R. R., Sulochana, K. G., & Sajini, T. (2011). Optimized Multi Unit Speech Database for High Quality FESTIVAL TTS. In Information Systems for Indian Languages (pp. 204-208). Springer Berlin Heidelberg.
27. Odell, J. J., Woodland, P. C., & Young, S. J. (1994, April). Tree-based state clustering for large vocabulary speech recognition. In Speech, Image Processing and Neural Networks, 1994. Proceedings, ISSIPNN'94, 1994 International Symposium on (pp. 690-693). IEEE.
28. Yamagishi, J. (2006). An introduction to hmm-based speech synthesis. Technical Report.
29. Masuko, T., Tokuda, K., Kobayashi, T., & Imai, S. (1997, April). Voice characteristics conversion for HMM-based speech synthesis system. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on (Vol. 3, pp. 1611-1614). IEEE.
30. http://sp-tk.sourceforge.net/- Speech signal processing tool kit (SPTK)
31. Digalakis, V. V., Rtischev, D., & Neumeyer, L. G. (1995). Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Transactions on speech and Audio Processing, 3(5), 357-366.
32. Gales, M. J., & Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. Computer Speech & Language, 10(4), 249-264.
33. Heiga, Z. E. N., Tokuda, K., Masuko, T., Kobayasih, T., & Kitamura, T. (2007). A hidden semi-Markov model-based speech synthesis system. IEICE Transactions on Information and Systems, 90(5), 825-834.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

34. Yamagishi, J., & Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE TRANSACTIONS on Information and Systems, 90(2), 533-543.
35. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., & Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR. adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing, 17(1), 66-83.
36. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039-1064.
37. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007, August). The HMM-based speech synthesis system (HTS) version 2.0. In SSW (pp. 294-299).
38. http://hts.sp.nitech.ac.jp/?Download
39. http://lib.tkk.fi/Dipl/2012/urn100698.pdf
40. Tokuda, K., Kobayashi, T., Masuko, T., & Imai, S. (1994, September). Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In ICSLP (Vol. 94, pp. 18-22).
41. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lecture%203_winter_2012_6tp.pdf
42. http://www.ijerd.com/paper/vol9-issue10/F09104854.pdf
43. http://www.phon.ucl.ac.uk/courses/spsci/iss/week1.php
44. Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv: 1003.4083.
45. Gopi, A., Shobana, P. D., Sajini, T., & Bhadran, V. K. (2013, December). Implementation of Malayalam text to speech using concatenative based TTS for android platform. In Control Communication and Computing (ICCC), 2013 International Conference on (pp. 184-189). IEEE.
46. Binil Kumar, S. L., Sajini, T., & Bhadran, V. K. (2013). Screen readers for Windows and Linux-Unit selection based Malayalam text to speech system integrated with disability aids Screen reader with Indian English.
47. Kominek, J., Black, A. W., & Ver, V. (2003). CMU ARCTIC databases for speech synthesis.
48. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Sixth European Conference on Speech Communication and Technology.
49. Stylianou, Y. (2009, April). Voice transformation: a survey. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on (pp. 3585-3588). IEEE.
50. Saheer, L., Garner, P. N., Dines, J., & Liang, H. (2010, March). VTLN adaptation for statistical speech synthesis. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 4838-4841). IEEE.
51. Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (1998). Speaker adaptation for HMM-based speech synthesis system using MLLR. In the third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.
52. Helander, E., & Gabbouj, M. (2012). Jani Nurminen1, Hanna Silén2, Victor Popa2. SPEECH ENHANCEMENT, MODELING AND RECOGNITION–ALGORITHMS AND APPLICATIONS, 69.
53. Zen, H., & Tokuda, K. (2009). TechWare: HMM-based speech synthesis resources [Best of the Web]. IEEE Signal Processing Magazine, 26(4).

Sajini T is currently pursuing MTech degree in Signal Processing with the Department of Electronics and Communication Engineering at College of Engineering Trivandrum, Kerala. She received her BTech degree from Cochin University of Science and Technology (CUSAT) in the year 2000. She is currently working as Principal Engineer, at Center for Advanced Computing (CDAC), Thiruvananthapuram. Her research area includes Speech signal processing, Speech synthesis, NLP, Deep neural network based synthesis and Brain signal processing.

Neetha George is an Assistant Professor in the Department of Electronics and Communication Engineering at College of Engineering, Thiruvananthapuram. She received her M.Tech. on Microelectronics & VLSI from NIT Calicut. She is currently doing research in Image Processing and her areas of interest include VLSI Design, Electronics Circuits, and VLSI for DSP.