2011 IEEE International Workshop on Machine Learning for Signal Processing
September 18-21, 2011, Beijing, China

IMPLEMENTATION AND EVALUATION OF STATISTICAL PARAMETRIC SPEECH SYNTHESIS METHODS FOR THE PERSIAN LANGUAGE

Sara Bahaadini, Hossein Sameti, and Soheil Khorram

Speech Processing Lab, Department of Computer Engineering, Sharif University of Technology,
Tehran, Iran
bahaadini@ce.sharif.edu, sameti@sharif.edu, khorram@ce.sharif.edu

ABSTRACT

Scattered and little research in the field of Persian speech synthesis systems has been performed during the last ten years. Comprehensive framework that properly implements and adapts statistical speech synthesis methods for Persian has not been conducted yet. In this paper, recent statistical parametric speech synthesis methods including CLUSTERGEN, traditional HMM-based speech synthesis and its STRAIGHT version, are implemented and adapted for Persian language. CCR test is carried out to compare these methods with each other and with unit selection method. Listeners Score samples based on CMOS. The methods were ranked by averaging the CCR scores. The results show that STRAIGHT-based system produces the best quality. Traditional HMM-based and unit selection are second and third in quality ranking. These approximately produce the same quality. Finally CLUSTERGEN produces the worst quality among these four systems.

Index Terms**—** text to speech, statistical parametric, speech synthesis, Persian language, CCR test

1. INTRODUCTION

The aim of this paper is to implement, adapt and evaluate Statistical parametric speech synthesis framework for Persian language. Traditional methods for speech synthesis are knowledge-based ones. These methods produce unnatural and machinery voices. As computers power increased corpus-based methods became more common. Unit selection [1] is one of the most successful examples of corpus-based speech synthesis methods. It is used in many commercial products. Despite its high prevalence, there are two major drawbacks with this method. The need for huge database to obtain acceptable quality and the low flexibility in modifying synthesized speech [2].

For synthesizing speech in this method, each required unit must be chosen form a prerecorded dataset. In order to obtain good quality, huge database with appropriate unit coverage must be collected [1]. Therefore, synthesized speech quality extremely depends on its database. Because of numerous possible combinations, there is no guarantee that training dataset contains all of the required phonetic and prosodic context units. When one of these uncovered contexts is met in the synthesis phase, unit selection algorithm fails to select the appropriate unit and this may ruin the listeners flow. This drawback turns unit selection method into a high-memory consumer method. Consequently it is impossible to employ unit selection method in low resource applications such as mobile handsets. Other major drawback of unit selection is its low flexibility; generating different intonations, styles and emotions in this method is severely difficult and inefficient [2].

Statistical approaches have recently been shown as very effective methods in various fields of speech processing. Statistical models, such as hidden Markov model (HMM), perform very well in speech recognition. Nowadays statistical parametric speech synthesis methods [2] are growing rapidly. HMM-based speech synthesis is implemented for many languages [3-9]. Statistical speech synthesis methods studied here for Persian language are traditional HMM-based speech synthesis [10], HMM-based with STRAIGHT vocoding [11] and CLUSTERGEN [12]. All the statistical parametric methods synthesize speech through the following steps

1- Extract suitable parameters (which will be needed for synthesizing speech) from the training utterances.
2- Model the parameters using one of the generative statistical methods.
3- Generate parameters from the trained models.
4- Synthesize speech according to the generated parameters.

Extracted parameter's behavior should be simple enough to be modeled easily in the subsequent steps. Furthermore they should contain as much information as needed so that generated speech from them is less distorted. In spectral parameters that are modeled for speech recognition systems, large amounts of information are discarded. Therefore the quality of synthesized speech by

978-1-4577-1623-2/11/$26.00 ©2011 IEEE

them in statistical methods is lower than the unit selection cases where recorded speech segments are concatenated and synthesized utterances are constructed from them. As statistical methods are always parametric ones, they are flexible; the generated speech could be modified extensively by changing the system parameter values. Speaker adaptation [13], interpolation [14] and eigenvoices [15] are three examples of this flexibility which enable the synthesized speech to be modified without the necessity of large datasets.

According to the reasons given for the advantages of statistical synthesizers, a comprehensive framework seems to be extremely necessary in Persian text to speech researches. Here three statistical methods are proposed for Persian language. A CCR test [16] is carried out to compare these methods with each other and with unit selection method. Listeners score the samples according to Comparison Mean Opinion Score (CMOS) measure [17].

This paper is organized as follows. In Section 2 the HMM-based technique is described. In Section 3 STRAIGHT vocoding version of the HMM-based method is explained. In Section 4, CLUSTERGEN method is described. Section 5 which is the most important part belongs to adaptation for Persian language. Experiments and results are described in Section6.

2. HMM-BASED SPEECH SYNTHESIS TECHNIQUE

In this method, source-filter model is used for speech production. Parameters that are necessary for producing speech are modeled by appropriate HMMs. In the training phase HMM models,$\hat{\lambda}$,are constructed as in Eqn. (1).

$$\hat{\lambda} = argmax_{\lambda}\{p(o|w,\lambda)\} \quad (1)$$

In speech synthesis phase, the required parameters are generated from trained models as Eqn. (2) and are dispatched to speech synthesis filter [2].

$$\hat{o} = argmax_{O}\left\{p\left(o|w,,\widehat{\hat{\lambda}}\right)\right\} \quad (2)$$

Source–filter model of speech production, the works done based on HMM-based method, and the activities carried out in order to adapt this system for Persian language are described in following sections.

2.1. Source–filter model of speech production

Speech synthesis methods that simulate human speech production mechanism are mainly based on the source/filter theory. In this model, speech is the result of passing a source signal (e(n)) through a filter (h(n)) as shown in Fig.1.The source signal simulates the voice generated by vibration of the vocal folds in the larynx and the filter models the vocal tract and modifies the source signal to generate phoneme sounds. A transfer function is used to formulate this filter. Final speech signal can be computed as Eqn. (3).

$$s(n) = h(n) * e(n) \quad (3)$$

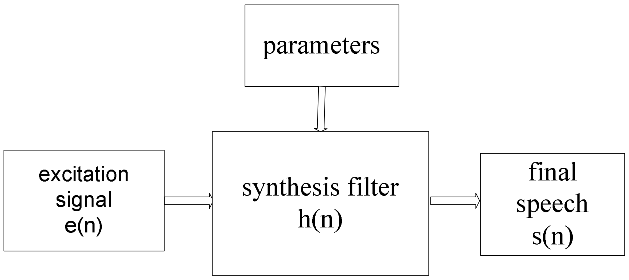


Figure 1: Source-Filter model

2.2. Speech modeling and production

Two categories of parameters are modeled: Spectral and Excitation. The mel-cepstral coefficients [18], their delta and delta-delta values are used for spectrum. The fundamental frequency consists of log F0 in addition to its delta and delta-delta for excitation. Spectrum features, F0 and state duration are modeled by a unified framework of context-dependent HMM. Multivariate Gaussian mixtures are used for spectral features. F0 is modeled by a multi space distribution [19]. The state durations of HMMs are modeled by a multivariate Gaussian distribution.

The overall system is shown in Fig.2. Initially mel-cepstral coefficients and F0 of the training data utterances are extracted. Then for each phoneme a context-independent HMM is trained as an initial model for the corresponding context-dependent model. Numbers of all contextual combinations are very high and the training dataset may not have sufficient training data for all of them, on the other hand in the synthesis part, a contextual combination not visited in the training data may occur. So the models of mel-cepstral, F0 and state duration should be clustered. In the next step, models are clustered using decision tree structures and similar models are tied based on predetermined criteria. The tied models are trained by more training data compared to the original context dependent models. Then models are untied and trained with their own training data. This procedure is done repeatedly until sufficient accuracy is achieved.

In the synthesis phase, each sentence is converted to the sequence of context-dependent labels. In our work this is done by constructing a hierarchical tree for sentence utterance structure. Contextual information are highly language dependent and more explanation on them are given in Section 5 which describes the adaptation aspects for the Persian language. In the next step of speech synthesis phase, parameter generation is done using context-dependent HMMs. The extracted mel-cepstral parameters are set as Mel Log Spectrum Approximation filter (MLSA filter) [20] parameters. F0 is used for excitation signal generation. A simple impulse train is used as the excitation signal. In the next section, mixed excitation signal based on STRAIGHT is studied. The excitation signal is filtered by the MLSA filter and final speech is produced.

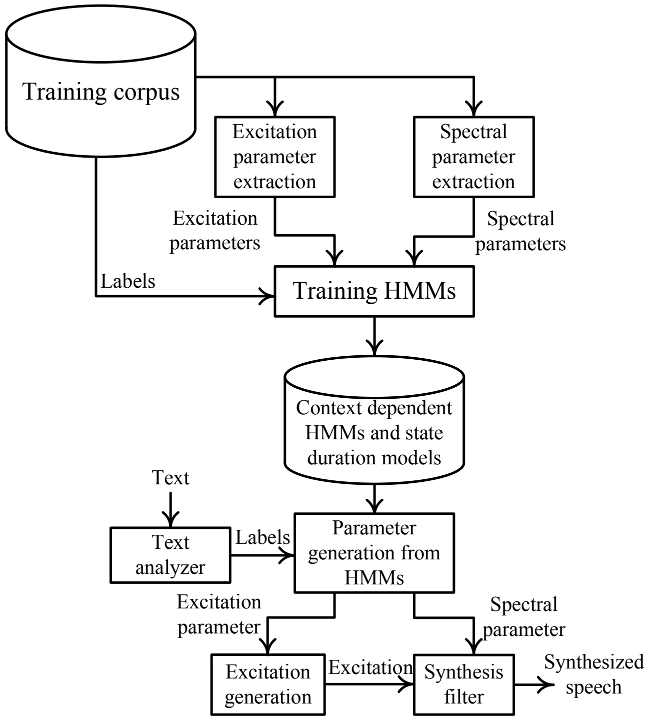


Figure 2: The overall HMM based speech synthesis system [2]

3. STRAIGHT BASED SYSTEM

STRAIGHT based system is an improvement over traditional HMM based speech synthesis. It uses a new vocoding algorithm [11]. This is done in three successive steps.

In the first step F0 is calculated by fixed-point analysis. Then the periodicity of signal is removed in time domain by F0-adaptive spectral analysis with surface reconstruction method [11]. Then methods for measuring aperiodicity of signal are applied. In the synthesis phase a mixed excitation signal is produced by a weighted mixture of white noise and pulse train. A phase manipulation is done on the excitation signal. Aperiodicity value is used in the weighting process in frequency domain. Final speech is constructed by smoothed spectrum and mixed excitation [11].

Here an HMM based system with STRIAGHT vocoding is implemented for Persian language and a comparison is done between the synthesized voices.

4. CLUSTERGEN

CLUSTERGEN is a simple and fast parametric synthesizer which requires well recorded utterances with their phonetic and prosodic transcriptions. In the training phase, corpus is labeled in the phonetic level through applying Baum Welch and training context independent HMM for each phoneme. Then a pitch synchronous frame size window is applied on training signals. For each frame an F0 and 24 MFCC coefficients are extracted. Then context information, required for clustering, is extracted for each frame. A CART tree is built for MFCC, F0 and duration independently for each HMM state.

In synthesis phase, an input text is converted to phoneme sequence. Duration CART trees of input phoneme HMM states are traced and corresponding values in their leaf are chosen. The average of these values is considered as each state duration. MFCC and F0 CART trees are also traced. Final speech is generated from these parameters by MLSA filter [20].

5. ADAPTATION FOR PERSIAN LANGUAGE

Adaption is done by considering special characteristics of Persian language including utterance structure and the contextual factors. In this study corpus design and implementation and Persian language analysis are done. In the following sections more details are described.

5.1. The Persian database for speech synthesis

This section gives an overview of the speech database used for generating various Persian synthesis systems. The database is designed after the Arctic database used in English. The classic Persian databases such as FARSDAT were collected to support speech recognition applications. Due to the requirements of recognition systems classical databases containing phonetically balanced utterances which read by various speakers are used. However in synthesis applications a single speaker database is required. Therefore here a new database with the goal of speech synthesis research is provided. The database comprises of two speaker utterances. These utterances are designed in order to satisfy following conditions:

-Each sentence is short enough to be recorded easily.

-The utterances are phonetically balanced. They properly cover Persian diphones and syllables.

- A professional speaker can read all sentences in a single day.

These sentences are selected from Peykare corpus [21]. This corpus contains 10 million words and 336,000 sentences. For the sake of recording it is reduced down to a list of 50,000 simple sentences. The selected sentences are between 5 and 20 words long and out of vocabulary words are avoided in them. Final sentences are selected according to the following steps.

Step1: 460 sentences to cover most frequent Persian words.
Step2: 40 sentences to cover all biletter combinations.
Step3: 100 sentences to cover all biphoneme combinations.
Step4: 545 sentences to cover most frequent Persian syllables.

5.2. Persian linguistic information

In this section, linguistic information, including important and effective information in Persian speech synthesis, is studied. Main structures of Persian language include Phonemes, syllables, words, phrases and sentences. Each of these has unique properties in Persian language that must be considered in text to speech procedure.

Phoneme

30 phonemes including 23 consonants, 6 vowels and a silence phoneme are considered. In this study Phone features considered are phoneme length, height, frontness, lip rounding. For consonants phoneme features such as consonant type (stop, fricative, affricative, nasal, liquid), place of articulation (labial, alveolar ...), consonant voicing are extracted.

Syllable

Persian language has important and very useful syllable structure. Syllables structures in Persian language are CV (type 1) or CVC (type 2) or CVCC (type 3). From position of vowel in word, syllables of word can easily be recognized.

Word

In this study three features are extracted for words. Stress pattern of word, Guess Part Of Speech (GPOS) tag and the third feature is “Ezafe”.

Stress is an important feature for pronouncing Persian words in a sentence and it can produce different meaning. One of the syllables in a word is uttered with more pressure. In Persian language usually the last syllable takes this pressure. When words are pronounced in a sentence, this pressure may be degraded or removed. However in this research words are considered as they are out of sentences.

Comprehensive research is done on GPOS tagging of Persian language. Totally 25 different GPOS tags are extracted for Persian words. The most popular methods for determining the POS tags are based on HMM. In this research, an HMM structure with 25 states is employed. Parts of Peykare [21] corpus is used as the train and the test data sets. 30% of this corpus is separated for testing and 70% for training. This is a sequence tagging problem which can be performed using the Viterbi algorithm. The accuracy of the implemented algorithm is approximately 96%.

The last feature is “Ezafe”. It is a special feature for Persian language. Two other language Kurdish and Zazaki which are derived from Persian have “Ezafe” too. “Ezafe” is the short vowel “e” that is placed between two adjacent words. It is not written but is pronounced.

“Ezafe” is used in the following conditions.

- Between two nouns for demonstrating possession.
- Between nouns and adjectives for demonstrating associations.

It should be noted that the detection of “Ezafe” is an important and difficult problem.

5.3. Contextual factors

In this study contextual factors that affect reading style in Persian language are considered. What is uttered in training data is converted into complete contextual phoneme labels as well as the input text. The contextual factors that are taken into account are as follow:

– Phonetic level

- The phoneme before the previous phoneme, the previous phoneme, the current phoneme, the next phoneme, the phoneme after the next phoneme.
- The position of the current phoneme identity in the current syllable(forward and backward).
- Whether this phoneme is “Ezafe” or not

– Syllable level

For the previous, current and the next syllable the following factors are considered:

- Whether it is stressed or not
- The number of phonemes
- The position in the current word and phrase (forward and backward)
- Type of Syllable (type 1 or 2 or 3)

And

- The number of stressed syllables before and after the current syllable in the current phrase
- The number of syllables from the previous stressed syllable to the current syllable
- The number of syllables from the current syllable to the next stressed syllable
- The vowel of the current syllable

– Word level

For the previous, current and the next word the following factors are considered:

- GPOS (guess part-of-speech) of word
- The number of syllables in word
- Position of the word in the current phrase (forward and backward)

And

- Whether this word is the last word in the sentence

– Phrase level

- The number of syllables in the current, next and previous phrase
- The position of the phrase in the sentence (forward and backward)

– Sentence level

- The number of phrases, words and syllables in the sentence

The described contextual information works as features. In the clustering step these features are used as answers to some of questions for decision tree construction. For each sentence a hierarchical structure (i.e., tree) is built based on the levels that are listed above. The first level is sentence; the second is phrase and so on. Information described in Section 5.2 is embedded into the data structure of the units of tree.

6. EXPERIMENTS AND RESULTS

Four previously described systems are evaluated here.

1. Traditional HMM based text to speech system with the simple impulse train as the excitation signal and 24 mel-cepstral coefficients.
2. STRAIGHT-based systems with STRAIGHT vocoding and with 39 mel-cepstral coefficients which are achieved from straight spectra [11].
3. CLUSTERGEN with 24 MFCC and F0 as the feature vector, 3 state HMMs, 5ms frame shift, pitch synchronous framing
4. Clustered Unit selection (CLU) which used optimal coupling technique, 12 MFCC coefficients without F0, the same weight for all elements of feature vectors in joint cost computation, the joint cost weight is 0.5 and the target cost is 1.

Subjective comparative tests were done between these 4 systems. Altered Comparison Category Rating (CCR [16]) tests were done between each pair of the synthesized speeches. In CCR, the qualities of the pair of outputs from two different systems are scored by listeners based on the 7-point Comparison Mean Opinion Score (CMOS) scale [17] .Table 1 shows CMOS scale.

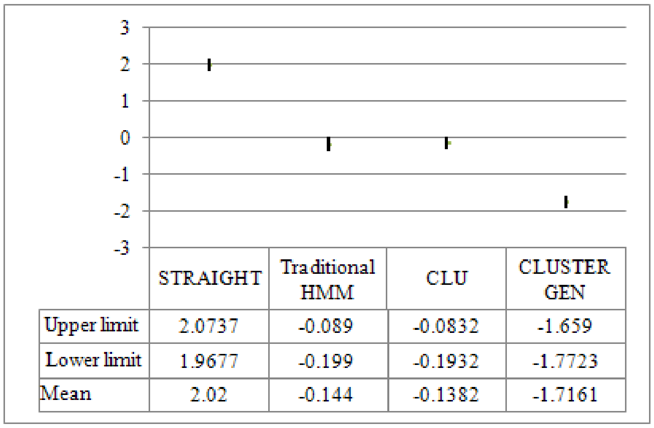
Table 1: CMOS Scores

Much better	+3
Better	+2
Slightly better	+1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

Totally 24 male and female listeners between 20 to 30 years of age did the test. The test samples included 20 fixed sentences generated by each of the traditional HMM-based, the STRAIGHT-based system, the CLUSTERGEN and the unit selection systems. All sentences were generated in random order by two randomly selected systems and played for the listeners. Listeners could play samples as many times as they wanted. They chose the better one and specified their level of preference according to CMOS scores.

The methods were ranked by averaging the CCR scores. 95% confidence intervals based on the 1-sided t-test is calculated in Eqn. (4).

$$upper\ limit = avg\ cmos_{test} + \frac{t_{N-1,\frac{\propto}{2}} * s\_cmos_{\,test}}{\sqrt{N}}$$
$$lower\ limit = avg\ cmos_{test} - \frac{t_{N-1,\propto/2} * s\_cmos_{\,test}}{\sqrt{N}} \qquad (4)$$



	STRAIGHT	Traditional HMM	CLU	CLUSTER GEN
Upper limit	2.0737	-0.089	-0.0832	-1.659
Lower limit	1.9677	-0.199	-0.1932	-1.7723
Mean	2.02	-0.144	-0.1382	-1.7161

Figure 3: Ranking of the CCR test for the proposed systems. The 95% confidence intervals are shown.

The final results are shown in Fig.3. The results show that STRAIGHT-based system produces the best quality followed by traditional HMM-based and unit selection .These approximately produce the same quality. Finally CLUSTERGEN produces the worst quality among these four systems.

7. CONCLUSION

A comprehensive framework for Perisan text to speech was needed for this field in Perisan language. In this work recent statistical parametric speech synthesis method were adopted for Perisan language. An appropriate Persian corpus for text to speech is designed and implemented. Persian linguistic information and contextual factors are considered in implementing text to speech methods. The synthesized voices were evaluated. The results shown that HMM-based with STRAIGHT vocoding produced the best results (around 2 CMOS score). Traditional HMM-based (around-0.14 CMOS score), unit selection (around -0.13 CMOS score) and CLUSTERGEN (around -1.7 CMOS score) are followed in order.

The systems quality will be used for future development of Persian language parametric text to speech.

8. ACKNOWLEDGMENT

This work was supported by the Iran Telecommunication Research Center.

9. REFRENCES

[1] A. Hunt, and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *In: Proc. of ICASSP*, Atlanta, Georgia, pp. 373–376, 1996

[2] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009

[3] S. Krstulovic, A. Hunecke, and M. Schroeder, ”An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements,” *In: Proc. of Interspeech*, Antwerp, 2007

[4] K. Tokuda, H. Zen, and A. Black, ”An HMM-based speech synthesis system applied to English,” *In: Proc. Of IEEE Speech Synthesis Workshop,* Santa Monica, California USA, 2002

[5] Y. Qian, F. Soong, Y. Chen, and M. Chu, “An HMM-based Mandarin Chinese text-to-speech system.,” *ISCSLP LNCS (LNAI)*, *Springer, Heidelberg,* vol. 4274, pp. 223–232. 2006

[6] R. Maia, H. Zen, K. Tokuda, T. Kitamura, F. Resende, ”Towards the development of a Brazilian Portuguese text-to-speech system based on HMM.,” *In: Proc. of Eurospeech*, Geneva, pp.2465–2468, 2003

[7]. X. Gonzalvo, C. Socoro, I. Iriondo, C. Monzo, E. Martı´nez, “Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish.,” *In: Proc. ISCA,* SSW6, pp. 362–367. 2007

[8] S.-J. Kim, J.-J. Kim, M.-S. Hahn, “HMM-based Korean speech synthesis system for handheld devices,” *IEEE Trans. Consumer Electronics* 52(4), 1384–1390 , 2006

[9] O. Abdel-Hamid, S. Abdou, and M. Rashwan, “Improving Arabic HMM based speech synthesisquality.,” *In: Proc. of Interspeech*, Pittsburg, pp. 1332–1335, 2006

[10] K. Tokuda, H. Zen, J. Yamagishi, A. Black, T. Masuko, S. Sako, T. Toda, T. Nose, K. Oura, and “The HMM-based speech synthesis system (HTS),” <http://hts.sp.nitech.ac.jp/>.

[11] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveign´e, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[12] A. Black, “CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling,” *In: Proc. Interspeech*, Pittsburgh, pp. 1762–1765, 2006

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speechsynthesis and a constrained SMAPLR adaptation algorithm,” *IEEETrans. Audio Speech Language Process.* 17 (1), pp. 66–83. 2009

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,“Speaker interpolation in HMM-based speech synthesis,” *In: Proc. of EUROSPEECH,* Lisboa, Portugal 1997

[15] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko,T. Kobayashi, and T. Kitamura, “Eigenvoices for HMM-basedspeech synthesis,” *In: Proc. of ICSLP*, Denver, Colorado, USA, 2002

[16] Recommendation ITU-U P.800, “Methods for subjective determination of transmission quality,” *In: International Telecommunication Union*, Aug. 1996.

[17] V. Grancharov, and W. Kleijn, “Speech Quality Assessment”, *Springer Handbook of Speech Processing,* chap. 5, 2007

[18] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” *In: Proc. of ICASSP*, Minneapolis, Minnesota, USA, vol.1, pp.137–140, 1992

[19] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. “Multi-space probability distribution HMM,”, *IEICE Trans*. Inf. Syst., E85-D(3):455–464, Mar. 2002

[20] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” *In: Proc. of ICASSP,* Boston, Massachusetts, USA, pp.93–96, Feb. 1983.

[21] M. Bijankhan, J. Seikhzadeghan, M. Bahrani, and M. Ghayoomi, "Lessons from Creation of a Persian Written Corpus: Peykare", *Language Resources and Evaluation Journal*, Springer Netherlands, vol.45, pp.143–164, 2010