Context-Dependent Deterministic Plus Stochastic Model

Soheil Khorram, Hossein Sameti, Fahimeh Bahmaninezhad
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
khorram@ce.sharif.edu, sameti@sharif.edu, bahmaninezhad@ce.sharif.edu

Abstract—**This article proposes a method to improve the performance of *deterministic plus stochastic model (DSM-)* based feature extraction by integrating the contextual information. One precious advantage of speech synthesis over speech recognition is that in both training and testing phases of synthesis, *contextual information* is available. However, similar to recognition, this invaluable knowledge has been forgotten during acoustic feature extraction of speech synthesis. DSM expresses the residual of Mel-cepstral analysis through a summation of two components, namely *deterministic* and *stochastic*. This study proposes to model the deterministic component through a novel *context-dependent principal component analysis (CD-PCA)*, and the stochastic component through the conventional high-pass filtered noise. Furthermore, due to the high dependency of the proposed feature extraction on state boundaries, the feature analysis and HMM-based modeling are performed in an iterative manner. Subjective evaluations conducted on a Persian speech database confirm the effectiveness of the proposed synthesis system.**

Keywords-context-dependent PCA, context-dependent residual modeling, excitation modeling, HMM-based speech synthesis, statistical parametric speech synthesis.

I. INTRODUCTION

Statistical parametric speech synthesis (SPSS) based on *hidden Markov model (HMM)* has dominated speech synthesis research area over the last decade [1]. It is a direct result of several favorable SPSS characteristics such as high flexibility to modify voice characteristics [2], capability to exploit all speech recognition techniques (e.g. adaptation methods) [2], proper support of multilingual synthesizers [3], improved coverage of acoustic space, low memory requirement [1].

All of the above advantages are achieved as consequences of statistical parametric representation of speech in SPSS. A typical SPSS system comprises two distinct phases [1], namely *training* and *synthesis*. Training phase starts with the extraction of acoustic features [4] and contextual factors [5] for all utterances in training database. Next, the relationship between acoustic features and contextual factors is modeled using a context-dependent statistical model [4]. In the synthesis phase, contextual factors are first obtained for an input text. Thereafter, a *parameter generation (PG)* algorithm [6] is employed to generate acoustic trajectories. Generated trajectories are then fed into a vocoder [7, 8] to generate synthesized speech.

However, this statistical parametric representation results in major quality reduction in synthesized speech [1]. This problem is known to be a result of three main issues, namely vocoding distortion [7, 8], deficiencies of statistical parametric models [1], and accuracy of parameter generation algorithms [6]. This paper is an attempt to alleviate the first issue and improve the performance of vocoders.

A. Related Work

Several efforts have been devoted to reduce the vocoding distortion in SPSS. A great number of them are based on the well-known source-filter model [9] initially inspired by human voice production system. This simple model generates speech by applying a certain filter to a source (excitation) signal. Physiologically, the excitation signal refers to the glottal air flow produced by vocal organs, and the filter indicates human vocal tract response. According to this physiological analogy, in order to have a reliable and efficient source-filter model, these glottal source excitation signal and vocal tract filter have to be separated from each other using glottal inverse filtering [9], which is an excessively difficult inverse problem. Therefore, several systems [4, 7, 8, 10-12] consider a much simpler framework instead of solving the difficult glottal inverse filtering problem. In their framework, the vocal tract filter captures the overall spectral envelope of speech and the glottal excitation signal corresponds to the residual signal obtained by passing the speech through the inverse of the estimated filter [10].

Traditionally, the widely-used *linear prediction (LP)* or *Mel-cepstral* coefficients are adopted to parameterize the spectral envelope of speech in the vocal tract filter. Furthermore, traditional systems make use of a random noise (for unvoiced frames) and an impulse train (for voiced frames) to synthesize the glottal excitation signal [1]. This naïve expression of the excitation signal is obviously not efficient enough, and synthesized speech using this excitation suffers from a strident *buzziness*. Accordingly, many research activities have been carried out to enhance the performance of the traditional excitation model. A large percentage of them rely on *mixed excitation (ME)* [4, 7, 8, 10-12] framework which generates the excitation signal through a superposition of both periodic and non-periodic components. Yoshimura et al. [4] was the first group that incorporated ME approach, used in MELP vocoder, in HMM-based speech synthesis. This system was later improved by Maia et al. [11] approach by applying two different state-dependent filters to the periodic and non-periodic components. The parameters of this model, including filter coefficients and the amplitudes of pulse train, are jointly optimized in the training phase of SPSS. *Liljencrants-fant (LF)* [10] glottal flow model is another system that has been shown to be reasonably effective for HMM-based speech synthesis. The LF model produces a waveform with a decaying spectrum at higher frequencies, which is more consistent with the natural source excita-

tion signal. *The speech representation and transformation using adaptive interpolation of weighted spectrum (STRAIGHT)* [12] is another popular ME-based vocoder used in parametric representation of speech. In this vocoder, aperiodicity measurements are defined to adjust the weight of periodic and non-periodic parts of excitation signal. It should be noted that STRAIGHT is currently considered to be one of the best vocoding methods for HMM-based speech synthesis.

Among all the above methods, *deterministic plus stochastic model (DSM)* [7, 8] is finally selected as the baseline system in this paper, due to its superior quality. DSM excitation is simply represented by a summation of two distinct components: a deterministic waveform, and a stochastic noise. Both components are trained using a speaker-dependent dataset of *pitch-synchronous (PS)* residual frames. In this approach, the noted *principal component analysis (PCA)* is responsible for expressing the deterministic part, and the stochastic part is synthesized through a high-frequency noise modulated both in time and frequency. Authors in [8] have reported that the DSM vocoder outperforms the traditional pulse excitation dramatically and provides a quality equivalent to STRAIGHT.

B. Scope of the Paper

The main idea of this study is to incorporate contextual information into the DSM vocoder in order to improve the performance of the predominant DSM. To this end, the PCA technique, applied to express the deterministic component of DSM, is replaced with a new method named *context-dependent PCA (CD-PCA)*. CD-PCA initially clusters all PS residual frames into several contextual groups. To perform the clustering, a greedy binary decision tree construction algorithm which minimizes the *root mean square error (RMSE)* of the generated residual signal is developed. This decision tree-based clustering scheme then contributes to generate the mean residual frame in CD-PCA. More precisely, in contrast to the conventional PCA which computes the mean component through a straightforward averaging, in the proposed CD-PCA, the mean component is generated through a weighted sum of many cluster prototypes; therefore, the mean vector obtained by CD-PCA is more similar to the target residual. Another important aspect of the proposed system is that its feature extraction and statistical modeling are mutually dependent, since on one side, feature extraction is a preliminary step for statistical modeling, and on the other side, changing statistical models leads to different state occupation probabilities and consequently different context dependent feature extraction. As a result, to have an optimum system, HMM training procedure and feature extraction have to be performed in an iterative manner.

The rest of the paper is organized as follows. In Section 2, the conventional DSM is briefly explained. Section 3 introduces the proposed system in detail. Experimental conditions and results are presented in Section 4 and final remarks are given in Section 5.

II. Deterministic Plus Stochastic Model

To explain the *context-dependent deterministic plus stochastic model (CD-DSM)*, first, a brief description of DSM is given in this section. DSM [7, 8] starts with the extraction of PS residual frames for all utterances in the dataset according to the following instructions. F0 trajectories and residual waveforms have to be extracted first. *Glottal closure instances (GCIs)* [13] are then recognized by locating the highest discontinuity in the residual signal. The desired PS frames are finally extracted using a GCI-centered with two period-long Blackman windowing.

After isolating the residual frames, deterministic and stochastic components of all frames have to be separated from each other. These components occupy two distinct spectral bands delimited by the maximum voiced frequency. Finally, the extracted deterministic and stochastic components are modeled independently.

A. Modeling of the Deterministic Component

As mentioned before, PCA approach is employed in DSM to decompose the low-frequency component on an orthonormal basis. Note that all frames have to be normalized in both duration and energy before applying the PCA. Two points have to be taken into account in length-normalization: 1) It should preserve the shape of all frames; 2) It should not lead to energy holes during synthesis process [8].

B. Modeling of the Stochastic Component

The method applied to represent the stochastic part is entirely in accordance with *harmonic noise model (HNM)* framework [14]. According to this framework, the stochastic waveform $r_s(t)$ is generated by convolving a white Gaussian noise $n(t)$ with an autoregressive model $h(\tau, t)$ and then controlling its time amplitude with a simple envelope $e(t)$ [14]. In other words,

$$r_s(t) = e(t).[h(\tau, t) * n(t)], \quad (1)$$

where $*$ denotes the convolution operator. Conventionally, the auto-regressive model $h(\tau, t)$ is considered to be identical for all frames of the dataset and therefore it does not require any parametric representation [8].

III. Context-Dependent DSM

DSM is a speaker-dependent vocoder; therefore, it is able to exploit speaker-specific information efficiently. However, contextual factors are not used during DSM procedures; while, they are available in speech synthesis applications. More precisely, DSM deals with various contexts in the same manner; while, it is possible to use different transformations for different contexts and improve the quality of conventional DSM.

Figure 1 compares three residual signals extracted from different phonemes of Persian language. As it is realized from this figure, these signals seem to be entirely different; therefore, designing the same residual modeling method for all contexts of the database is not a good option.

A. CD-DSM architecture

The overall architecture summarizing the proposed speech synthesis system that exploits the CD-DSM vocoder can be found in Figure 2. In accordance with SPSS framework, this system also consists of two phases: *train* and *synthesis*.

The train phase starts with feature extraction followed by context-dependent statistical modeling module. Features include mcep coefficients, F0 trajectory and contextual factors

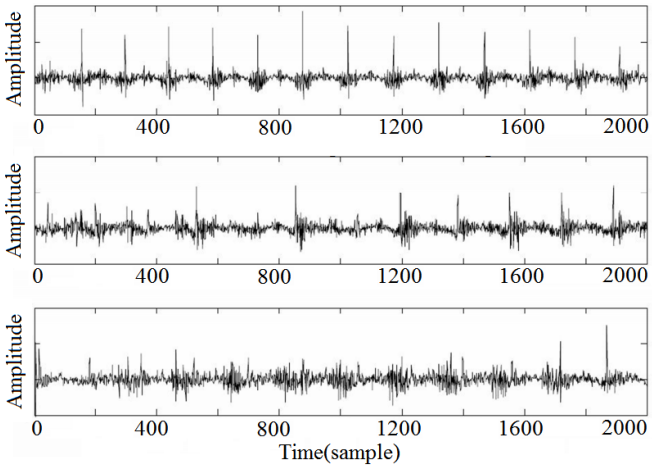


Figure 1. Residual signal extracted for phonemes /a/, /m/ and /@/.

that are incorporated with residual parameters. In order to extract the residual parameters, first, residual signal has to be obtained through applying spectral envelope inverse filter to the original speech waveform. The proposed CD-DSM analyzer is then employed to extract the desirable residual parameters. CD-DSM requires state occupation probabilities and contextual factors in addition to all inputs used in the conventional DSM (including residual signal and F0 trajectory). These occupation probabilities, on one side, are determined by stochastic modeling, and on the other side, modify the CD-DSM transform parameters applied in stochastic modeling. Therefore, CD-DSM and statistical modeling are mutually dependent and have to be trained together iteratively. Due to the intensive computation involved in each iteration, this paper proposes to initialize the occupation probabilities accurately in order to reduce the number of required iterations. An efficient initialization is to borrow the probabilities from a phoneme-dependent HMM simply trained with mcep stream. Using this initialization, the model can be converged rapidly (just in 2 or 3 iterations).

As it is shown in Figure 2, the synthesis phase is carried out by sequentially applying PG algorithm, CD-DSM residual synthesis and finally MLSA speech synthesis. To the extent that CD-DSM requires occupation probabilities, the PG algorithm has to be able to generate the probabilities along with acoustic trajectories. Certain PG algorithms, such as the *expectation maximization (EM-)* based method proposed in [15], implicitly compute the occupation probabilities, but many others, such as the *global variance (GV-)* based algorithm, cannot directly provide occupation probabilities. For the second group, it is possible to first generate acoustic trajectories, and then compute the probabilities by applying forward and backward algorithm [16] to generated trajectories.

B. CD-DSM analysis

A workflow explaining the CD-DSM analysis block is presented in Figure 3. According to this workflow, CD-DSM analysis is performed in 4 main steps as follows.

Step 1- PS framing: Similar to DSM, the first step is to isolate pitch-synchronous frames from the residual waveform. PS framing requires identifying the location of glottal closure instances in residual signal. *SEDREAMS* algorithm [13] is exploited for GCI detection in this study. PS frames are then simply extracted by applying a GCI-centered Blackman window of two pitch periods long. Figure 4 shows a typical residual signal and its extracted PS frame.

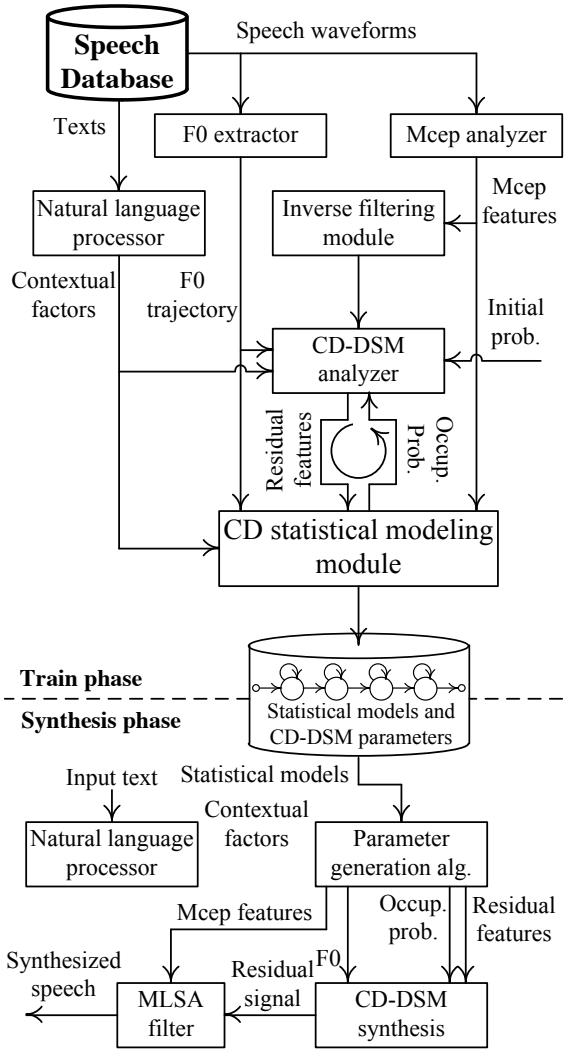


Figure 2. Block diagram of the proposed synthesis system with the embedded CD-DSM vocoder.

Step 2- DS separation: In this step, the residual signal has to be split in deterministic and stochastic components. It is assumed that these components lie in distinct spectral bands delimited by the maximum voiced frequency (4 kHz); therefore, DS separation can be accomplished using a simple hard-filtering method.

Step 3- Deterministic modeling: PCA is normally considered for deterministic part. Preliminary to PCA, PS frames have to be normalized both in pitch period and energy. Deterministic part is then decomposed into a number of orthonormal bases obtained by the proposed CD-PCA.

In CD-PCA, all normalized PS frames are clustered using a decision tree structure. The decision tree is a binary tree in which a contextual question is attached to each intermediate node and clusters are defined through terminal nodes. For each

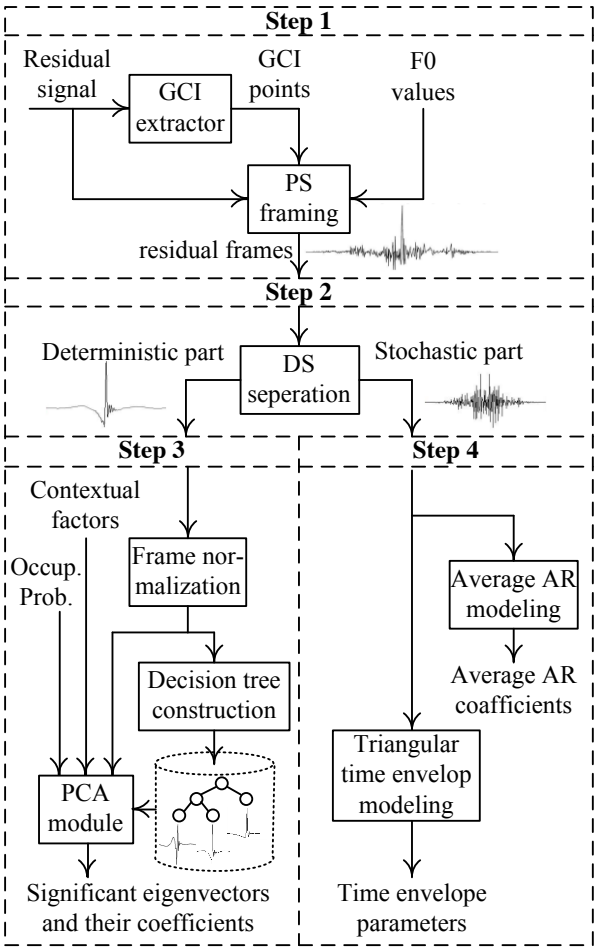


Figure 3. Workflow of the proposed CD-DSM analysis.

PS frame a single path from the root to one of the terminal nodes is traversed by recursively answering contextual questions and consequently the PS frame is assigned to a cluster.

The decision tree structure is built using a greedy top-down optimization procedure [16]. Initially, all of the PS frames are grouped into a unique cluster represented by the root node of the tree. A clustering error, which evaluates the efficiency of clusters, is also calculated in this step. The root node is then split into two nodes by finding the question which results in maximum reduction in clustering error. This procedure is then repeated by selecting the best pair of terminal node and question which yields the greatest decrease in clustering error until this decrease falls below a threshold. Generation error of PS residual frames is defined as the clustering error in this study.

To express the generation error mathematically, suppose we are given N deterministic residual frames $\{r_n(t)\}_{n=1}^{N}$, all normalized in T samples and clustered through a decision tree with C leaves. Additionally, let us define S as the total number of untied states in HMM, $\gamma_s(n)$ as the probability of occupying s-th state at n-th residual frame, and $I_c(s)$ as an indicator function of cluster c. $I_c(s)$ represents a binary function that takes the decision tree into account and determines whether the s-th state belongs to cluster c or not. According to the above notations, prototype frame of the c-th cluster, $\hat{r}_c(t)$, can be simply calculated by averaging all frames placed in that cluster. In other words,

$$\forall\, 1 \le t \le T \;\; \hat{r}_c(t) = \frac{\sum_{s=1}^{S} I_c(s) \sum_{n=1}^{N} \gamma_s(n) r_n(t)}{\sum_{s=1}^{S} I_c(s) \sum_{n=1}^{N} \gamma_s(n)}. \quad (2)$$

This prototype signal is the most reasonable signal that can be generated for each cluster; therefore, the generation error can be computed as:

$$E = \frac{\sum_{n=1}^{N} \sum_{s=1}^{S} \gamma_s(n) \sum_{c=1}^{C} I_c(s) \sum_{t=1}^{T} \left(r_n(t) - \hat{r}_c(t)\right)^2}{T \sum_{n=1}^{N} \sum_{s=1}^{S} \gamma_s(n) \sum_{c=1}^{C} I_c(s)}. \quad (3)$$

Using this error measure, the decision tree construction procedure can be accomplished. The decision tree structure and all cluster prototype signals obtained through Eq. (2) have to be saved for the next PCA module. Note that, in this study, 5 independent decision trees are trained for 5 states of HMM.

The residual frames are then decomposed into their principle components using CD-PCA. CD-PCA is slightly different from conventional PCA in such a way that local mean signals of each cluster (the prototype PS frame) is used instead of the conventional global mean. Localizing mean component of PCA using contextual information results in more accurate residual modeling and consequently more favorable synthesis system.

Step 4- Stochastic modeling: Stochastic modeling is completely identical to the noise component modeling in HNM [14]. It corresponds to a white noise modulated in both time and frequency as it is described in Section 2.2.

C. CD-DSM synthesis

The synthesis part of CD-DSM is simply accomplished through the reverse procedure of CD-DSM analysis. More specifically, deterministic and stochastic components are first constructed and then combined using overlap-and-add method to generate the output residual signal. It should be noted that mean component of CD-PCA is achieved by traversing decision trees and applying Eq. (2). Figure 5 shows the overall structure of the CD-PCA synthesis.

IV. EXPERIMENTS

We compare three speech synthesis systems based on traditional, DSM and CD-DSM excitation modeling approaches. Experimental conditions are first explained and then the results

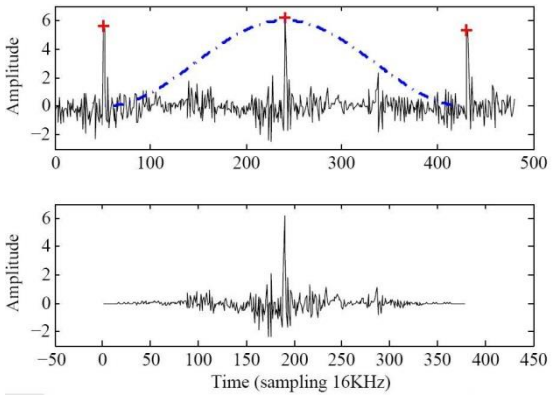


Figure 4. An example of the applied PS framing.

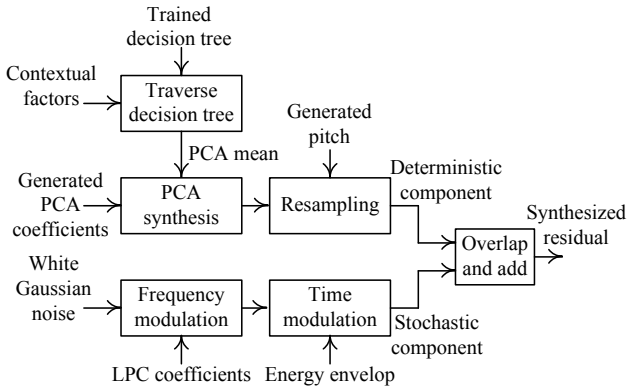


Figure 5. Block diagram of CD-PCA synthesis.

of subjective evaluations are reported.

A. *Experimental conditions*

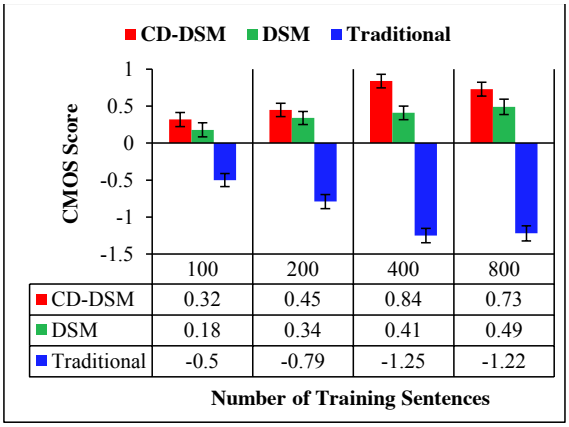
In order to train systems, a Persian speech dataset [17, 18] comprising 1000 sentences from a male speaker is employed throughout our experiments. Utterances of this database are between five to twenty words long and have an average duration of eight seconds. This database is specifically designed for the purpose of single-speaker speech synthesis applications. It covers most frequent Persian words, all bi-letter combinations, all bi-phoneme combinations, and the most frequent Persian syllables. 31 synthesis units, including 30 phonemes and a silence, are considered in the modeling phase. Speech waveforms are recorded with 16 kHz sampling rate and are windowed by a 25-ms Blackman window with a 5-ms shift. 40 Mel-cepstral coefficients, a fundamental frequency, residual features and their delta and delta-delta coefficients are employed as our acoustic features. Additionally, global variance (GV)-based parameter generation algorithm [6] is applied in the synthesis phase.

In our experiments, a multi-stream left-to-right with no skip path MSD-HMM [1] is trained as the acoustic model. Decision trees of context-dependent HMM-based acoustic models are built using maximum likelihood criterion and the sizes of trees are determined by MDL principle [1]. Publicly available HTS toolkit is slightly modified to be able to perform the HMM-based acoustic modeling phase. HTS with five streams of data are considered: one stream for the Mel-cepstral coefficients, one for the fundamental frequency, one for the derivatives of fundamental frequency, one for the PCA weights of the deterministic part, and one for the PCA weights derivatives.

All experiments are conducted on 4 different training sets including 100, 200, 400, and 800 utterances. Furthermore, a fixed set of 200 utterances, not included in the training sets, is employed as the test set.

B. *Experimental results*

Two well-known subjective evaluations are carried out in order to prove the effectiveness of the proposed system in databases with different sizes. These evaluations include a comparative mean opinion score (CMOS) test [7] with a 7-point scale, ranging from -3 (meaning that method A is much better than method B) to 3 (meaning the opposite), and a preference



	100	200	400	800
CD-DSM	0.32	0.45	0.84	0.73
DSM	0.18	0.34	0.41	0.49
Traditional	-0.5	-0.79	-1.25	-1.22

Figure 6. The result of the subjective evaluation using CMOS measure with 95% confidence intervals.

scoring [7]. Accordingly, twenty non-professional native listeners were presented with 30 randomly chosen pairs of synthesized speech generated by different systems. Listeners were asked to select the synthesized speech which sounds better and determine how much is better (much better, better, slightly better, or about the same). The results are shown in Figures 6 and 7.

Both CMOS tests and preference scores confirm the superiority of the proposed CD-DSM over DSM and traditional residual modeling approaches. As it can be seen in the figures, this superiority of CD-DSM is reduced for limited training databases. It is mainly due to the fact that statistical modeling is not accurate enough for small databases.

V. CONCLUSION

One of the main problems of *statistical parametric speech synthesis (SPSS)* systems, namely the distortion of vocoding procedure, was addressed in this paper. To alleviate this problem, it is proposed taking advantage of the contextual information in SPSS vocoding module. A novel *context-dependent principle component analysis (CD-PCA)* is designed to improve the performance of *deterministic plus stochastic model (DSM)* applied in excitation modeling of mcep-based vocoder. Conducted subjective tests confirm the improvement of the proposed method.

VI. REFERENCES

[1] Zen, H., Tokuda, K., and Black, A. W., "Statistical parametric speech synthesis", Speech Communication, Vol. 51, No. 11, pp. 1039-1064, 2009.

[2] Yamagishi, J., and Kobayashi, T., "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", transactions on information and systems (IEICE), Vol. 90, No. 2, pp. 533-543, April 2007.

[3] Gibson, M., Hirsimaki, T., Karhila, R., Kurimo, M., and Byrne, W., "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction", IEEE acoustics speech and signal processing (ICASSP), pp. 4642-4645, March 2010.

[4] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura. T., "Mixed excitation for HMM-based speech synthesis", European conference on speech communication and technology (INTERSPEECH), pp. 2263-2266, September 2001.

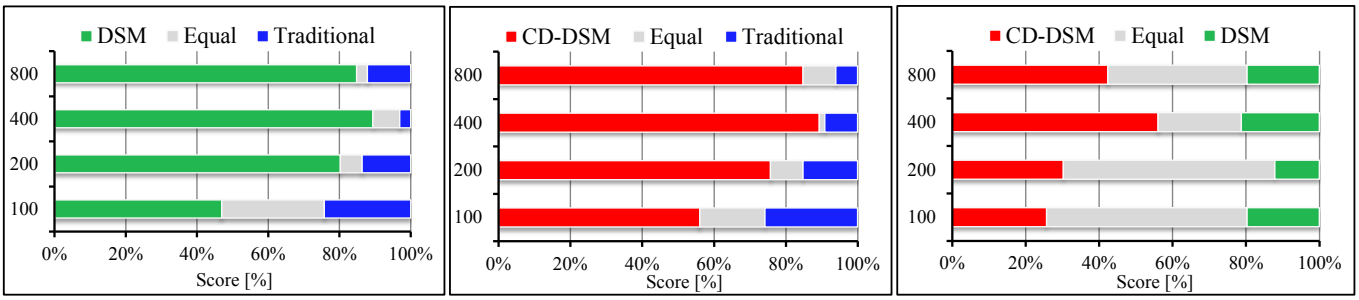


Figure 7. Preference score as a function of the number of utterances used for training: Comparison between (a) DSM and traditional, (b) CD-DSM and traditional, (c) CD-DSM and DSM.

[5] Sproat, R. "Multilingual text analysis for text-to-speech synthesis", International conference on spoken language processing (ICSLP), Vol. 3, pp.75-80, 1996.

[6] Toda, T., and Tokuda, K., "Speech parameter generation algorithm considering global variance for HMM-Based Speech Synthesis", Transactions on information and systems (IEICE), Vol. E90-D, No. 5, pp. 816-824, 2007.

[7] Drugman, T., Dutoit, T., "The deterministic plus stochastic model of the residual signal and its applications", IEEE transactions on audio, speech and language processing, Vol. 20, No. 3, pp. 968-981, March 2012.

[8] Drugman, T., Wilfart, G., and Dutoit, T., "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis", INTERSPEECH, pp. 1779-1782, Sep. 2009.

[9] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE transactions on audio, speech, and language processing, Vol. 19, No. 1, pp. 153-165, 2011.

[10] Cabral, J. P., "HMM-based speech synthesis using an acoustic glottal source model", PhD thesis, Edinburgh University, 2011.

[11] Maia, R., Toda, T., Zen, H., Nankaku Y., Tokuda K., "An excitation model for HMM-based speech synthesis based on residual modeling", speech synthesis workshop, pp. 131-136, 2007.

[12] Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech communication, Vol. 27, No. 3, pp. 187-207, 1999.

[13] Drugman, T. and Dutoit, T., "Glottal closure and opening instant detection from speech signals", INTERSPEECH, pp. 2891-2894, 2009.

[14] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis". IEEE transactions on speech and audio processing, Vol. 9, No. 1, pp. 21-29, 2001.

[15] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi T., and Kitamura T., "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", ICASSP'2000, Vol.3, pp. 1315- 1318, Istanbul, June, 2000.

[16] Odell, J., "The use of context in large vocabulary speech recognition", PhD dissertation, Cambridge University, 1995.

[17] Bijankhan, M., Sheikhzadegan, J., Roohani, M. R., Samareh, Y., Lucas, C., and Tebiani, M., "The Speech Database of Farsi Spoken Language", Proceedings of 5th Australian International Conference on Speech Science and Technology (SST'94), pp. 826-831, 1994.

[18] Khorram, S., Sameti, H., Bahmaninezhad, F., King, S., and Drugman, T., "Context-dependent acoustic modeling based on hidden maximum entropy model for statistical parametric speech synthesis", EURASIP Journal on Audio, Speech, and Music Processing, Vol. 1, No. 12, 2014.