A KL DIVERGENCE AND DNN APPROACH TO CROSS-LINGUAL TTS

Feng-Long Xie[1,2*] *Frank K. Soong*[2] *Haifeng Li*[1]

[1]Harbin Institute of Technology, Harbin, China
[2]Microsoft Research Asia, Beijing, China
{v-fxie,frankkps}@microsoft.com, lihaifeng@hit.edu.cn

ABSTRACT

We propose a Kullback-Leibler divergence (KLD) and deep neural net (DNN) based approach to cross-lingual TTS (CL-TTS) training. A speaker independent DNN (SI-DNN) ASR is used to equalize the speaker difference between a source speaker in L1 and a reference speaker in L2. Two speaker dependent GMM-HMM parametric TTS systems are first trained in the respective languages. The senones sets of the two TTS are matched in the SI-DNN ASR in terms of their output posteriors distributions in KLD. The minimum KLD criterion is used to transform the senones in the source speaker's TTS (L1) to the corresponding "closest" senones in the target language (L2). The new CL-TTS thus trained has been shown to achieve high speaker similarity to the source speaker in L1 while high intelligibility and naturalness are preserved. For untranscribed source speaker's recordings, say, conversational speech, a frame mapping, instead of "senone mapping" is also proposed to achieve a high but slightly inferior CL-TTS.

Index Terms— cross-lingual, speech synthesis, Kullback-Leibler divergence, deep neural networks

1. INTRODUCTION

Cross-lingual TTS synthesis is to synthesize speech in the target language (L2) with a specific speaker's recorded speech in source language (L1) and to maintain this speaker's voice characteristics, i.e., timbre. It has many applications, e.g., in speech-to-speech translation, it is highly preferable that the translated phrase can be synthesized in speech similar to the source speaker's voice. In 2nd language learning, a cross-lingual TTS system with the learner's own voice timbre can be useful and motivating to a learner. Several approaches have been proposed, including: GMM-HMM TTS state mapping[1] trajectory tiling[2] and spectral space warping[3]. These 3 approaches have achieved reasonably good performance. In [1] two separate, language-specific decision trees are built with Mandarin and English speech data recorded by a bilingual reference speaker, and the terminal leaves of the decision tree in L2 are mapped to the corresponding nearest neighbored terminal leaves of the the decision tree in L1. A new CL-TTS in L2 can then be built with the data recorded by the source speaker in L1. However, it's usually difficult to find a professional bilingual speaker. In [2], a reference speaker in the target language (L2) is used to help building the target language TTS with "tiles" of the original source speaker's monolingual (L1) data. This method can achieve highly intelligible synthesized speech, the similarity to the original source speaker can be further improved due to the fact that the speakers' difference is only equalized by a single parameter based bilinear warping function. In unit selection based speech synthesis, phone mapping can be used to find suitable units from the speech data of a monolingual speaker in the source language to synthesize speech in the target language. Such a mapping, in general, is based upon a good acoustic-phonetic similarity measure and mapping between two different languages at a phonemic level is almost by definition imperfect since the phonemic spaces of the two languages are not identical [4]. In [5], speech waveform segments from the speaker's source language database are selected when their acoustic cepstrum are similar to the reference speaker's segments in the target language. In the past few years, Deep Neural Networks (DNN) has been successfully applied to speech recognition[6]. Context-dependent, deep-neural-network HMMs (CD-DNN-HMMs), apply the classical ANN-HMMs of the 90's to the traditional tied-state triphones directly, by exploiting the pre-training procedure[7]. DNN architectures generate compositional models, where extra layers can enable composition of features from lower layers, giving them a huge learning capacity to model complex patterns of speech data. The long window of frames in DNN input also can incorporate more temporal information in a longer context.

In this paper, a speaker independent, deep neural network (SI-DNN) ASR is trained and the corresponding ASR senones space i.e., clustered GMM states, are used to represent the whole phonetic space speaker independently. Speaker differences can then be equalized with the SI-DNN at the senone or frame level in the phonetic space. KLD [15] is used to measure the difference between the source speaker's L1 senones and the reference speaker's L2 senones after the two speakers' difference is equalized. Thus the senone mapping is established based on a minimum KLD criterion. Finally the source speaker's L2 GMM-HMM TTS can be constructed with the senone mapping result. We also propose frame mapping when the transcriptions of the source speaker's L1 speech are not available.

The rest of this paper is organized as follows. In section 2 we will briefly introduce symmetrised KL Divergence used in this study. In section 3 the framework of the KLD-DNN approach to CL-TTS is proposed. In section 4 we describe experiments used to evaluate the performance of the proposed method. Finally we given our conclusions in section 5.

2. SYMMETRISED KL DIVERGENCE

The Kullback-Leibler divergence [15] (also known as information divergence, information gain, relative entropy, etc.) is a non-symmetric measure of the difference between two probability distributions, P and Q, which can be measured in a discrete or a continuous density form. For discrete probability distributions P

*Work performed as an intern in the Speech Group, Microsoft Research Asia

and Q, the KL divergence of Q from P is defined as,

$$\boldsymbol{D}_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{1}$$

For distributions P and Q of continuous random variables, the KL divergence is defined as an integral,

$$\boldsymbol{D}_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{2}$$

where p and q denote the probability densities of P and Q.

In this study we use a symmetrised discrete form of KLD defined as ,

$$\begin{aligned} \boldsymbol{D}_{KL}(P,Q) =& \boldsymbol{D}_{KL}(P||Q) + \boldsymbol{D}_{KL}(Q||P) \\ =& \sum_i (P(i) - Q(i))(\ln(P(i)) - \ln(Q(i))) \end{aligned} \tag{3}$$

It's positive semi-definite (i.e., $\geqslant$0), and we use it to measure the distortion between two given discrete distributions which are computed as the DNN output posterior probabilities.

3. KL DIVERGENCE AND DNN APPROACH TO CROSS-LINGUAL TTS SYNTHESIS

In statistical parametric speech synthesis, we construct a decision tree by clustering context-dependent, hidden Markov model (HMM) states to represent the probability densities of speech parameters of a given "senone"(context-dependent clustered state). In this section we use a speaker independent DNN to equalize the difference between the source speaker and the reference speaker in different languages. The SI-DNN trained in L1 data can find a proper posterior distribution for each TTS (L2) senone in a cross-lingual scenario, under the assumption that different spoken languages share the same acoustic-phonetic space at the senone, i.e., a sub-phonemic level. Since the output posterior distribution of SI-DNN is speaker independent, we can use KLD to measure the phonetic distortions between L1 and L2 in the probability space. The TTS (L1) senones of the source speaker can then be mapped to the closet senone in the reference speaker's TTS (L2) to create a cross-lingual TTS for the source speaker in L2.

3.1. Senone Mapping

A block diagram of senone mapping in KLD-DNN based CL-TTS training is shown in Fig. 1.

In training, both source speaker's L1 speech and reference speaker's L2 speech are collected. The two speech corpora are first used to train two GMM-HMM based TTS systems in the corresponding languages, respectively. Given the senone-level forced alignments of L1 training speech, we can get I buckets of training data, where I is the number of TTS (L1) senones. Then we process each bucket of source speaker's training data in L1 via the SI-DNN to get the accumulated posteriors for all N English ASR senones. The accumulated posteriors are then averaged by the number of accumulated frames in each bucket. For each TTS senone in L1 s_i^{L1}, we obtain the ASR senone posterior distribution P_i. In the same way, we obtain ASR senone posterior distribution Q_j for each TTS senone in L2 s_j^{L2}.

$$\begin{gathered} P_i = [\ p(s_1^{ASR}|s_i^{L1})\ p(s_2^{ASR}|s_i^{L1})\ \ldots\ p(s_N^{ASR}|s_i^{L1})\ ], \\ i \in [1,2,...,I] \end{gathered} \tag{4}$$

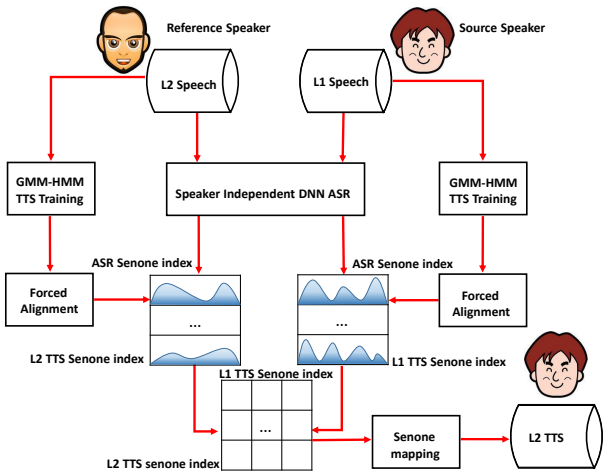

Fig. 1. KLD-DNN Cross-Lingual TTS: senone mapping

$$\begin{gathered} Q_j = [\ q(s_1^{ASR}|s_j^{L2})\ q(s_2^{ASR}|s_j^{L2})\ \ldots\ q(s_N^{ASR}|s_j^{L2})\ ], \\ j \in [1,2,...,J] \end{gathered} \tag{5}$$

We use KLD to measure phonetic distortion between each TTS (L1) senone s_i^{L1} distribution P_i and each TTS (L2) senone s_j^{L2} distribution Q_j in the probability space. A senone mapping is established with the minimum KLD selection criterion. For each TTS (L2) senone s_i^{L1} of the reference speaker, we find a corresponding TTS (L1) senone $s_{Map(i)}^{L2}$ of the source speaker in the minimum KLD sense.

$$\begin{aligned} Map(i) =& \operatorname*{argmin}_j \boldsymbol{D}_{KL}(P_i, Q_j) \\ =& \operatorname*{argmin}_j \sum_{n=1}^{N} (p(s_n^{ASR}|s_i^{L1}) - p(s_n^{ASR}|s_j^{L2}))* \\ & (\ln(p(s_n^{ASR}|s_i^{L1})) - \ln(p(s_n^{ASR}|s_j^{L2}))), \\ & i \in [1,2,...,I],\ \ j \in [1,2,...,J] \end{aligned} \tag{6}$$

Finally the source speaker's L2 TTS can be constructed with his own L1 TTS senones after the minimum KLD matching. In our HMM-based speech synthesis, spectrum, pitch and duration features are separated into three streams and three separate stream-dependent decision trees are built to cluster context-dependent states. The senone mapping is only established for spectrum matching.

3.2. Frame Mapping

In certain scenarios, L1 speech of the source speaker may be collected conversationally without any prescribed text. To deal with such a situation, we propose a frame mapping based KLD-DNN derived CL-TTS as shown in Fig. 2. For each L2 TTS senone s_j^{L2} we can find an ASR senone posterior distribution Q_j as in Eq. (5).

Given the source speaker's utterances in L1 without transcriptions, we can similarly get a posterior distribution P_k across all the ASR N senones in the SI-DNN for the k-th frame x_k.

$$\begin{gathered} P_k = [\ p(s_1^{ASR}|x_k)\ p(s_2^{ASR}|x_k)\ \ldots\ p(s_N^{ASR}|x_k)\ ] \\ k \in [1,2,...,K] \end{gathered} \tag{7}$$

We can then measure the phonetic distortion between the posterior distribution P_k of each source speaker's L1 speech frame, x_k, and

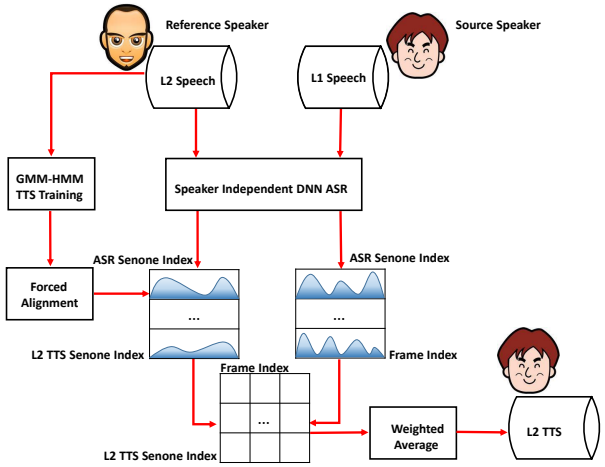

Fig. 2. KLD-DNN Cross-Lingual TTS: frame mapping

the posterior distribution Q_j of each TTS (L2) senone s_j^{L2}. For the k-th frame x_k, we obtain J such KLDs to J different TTS (L2) senones. They can be used as a weight for each TTS (L2) senone as in Eq.(9). The acoustic mean μ_j and variance σ_j^2 of each TTS (L2) senone s_j^{L2} for the source speaker can then be computed in a weighted average as

$$\begin{aligned} \boldsymbol{D}_{KL}(x_k, s_j^{L2}) = \sum_{n=1}^{N} (p(s_n^{ASR}|x_k) - p(s_n^{ASR}|s_j^{L2}))* \\ (\ln(p(s_n^{ASR}|x_k)) - \ln(p(s_n^{ASR}|s_j^{L2}))) \end{aligned} \tag{8}$$

$$w(s_j^{L2}|x_k) = \frac{e^{-\boldsymbol{D}_{KL}(x_k, s_j^{L2})}}{\sum_{j=1}^{J} e^{-\boldsymbol{D}_{KL}(x_k, s_j^{L2})}} \tag{9}$$

$$A_j = \sum_{k=1}^{K} w(s_j^{L2}|x_k) \tag{10}$$

$$\mu_j = \frac{1}{A_j} \sum_{k=1}^{K} w(s_j^{L2}|x_k) x_k \tag{11}$$

$$\sigma_j^2 = \frac{1}{A_j} \sum_{k=1}^{K} w(s_j^{L2}|x_k)(x_k - \mu_j)^2 \tag{12}$$

where K is the accumulated number of fractional frames.

3.3. Prosody Transformation

In this study we concentrate on preserving the timbre, i.e., speaker's voice characteristics in the spectral domain. However, the prosody of the source speaker is still transformed but only in a global scale between the reference speaker (L2) and the source speaker (L1). In synthesis , we generate the reference speaker's pitch trajectory first, and then a Gaussian normalized transformation[11] is used to transform the $F0$ of the reference speaker (L2) to the $F0$ of the source speaker (L1) as follows:

$$\ln(F0_{Trans}) = \mu_{L1} + \frac{\sigma_{L1}}{\sigma_{L2}}(\ln(F0_{L2}) - \mu_{L2}) \tag{13}$$

where μ and σ are the means and standard deviations of the two speakers in the corresponding languages.

The source speaker's L2 TTS duration model is directly copied from the reference speaker.

4. EXPERIMENTS

We evaluate the intelligibility, naturalness and similarity of the speech synthesized by our KLD-DNN based, cross-lingual TTS system. English and Mandarin belong to two different language families and they have many significant differences which makes English-Mandarin cross-lingual TTS synthesis more challenging. In this study, we take English as L1 and Mandarin as L2. A male, bilingual speaker with a mother tongue of Taiwanese Mandarin and a learned second language English is used as the source speaker. Only his English speech is used to synthesize his Mandarin speech. His Mandarin TTS trained with his Mandarin speech is used as a benchmark of the upper bound performance of our cross-lingual TTS experiments. A native male Mandarin speaker M is adopted as the reference speaker.

4.1. Experimental Setup

A database of 1,000 Mandarin utterances ($\sim$ 1 hour) of a reference native male speaker is used for training a speaker dependent, Mandarin GMM-HMM based TTS. Speech is sampled at 16kHz, windowed by a 25ms windows, and shifted every 5ms. 40th-order Line Spectral Pair (LSP) coefficients [9] plus gain and corresponding first and second order dynamic features, the fundamental frequency(F0) in log scale and its first and second order dynamic features are extracted. Multi-space probability Distribution (MSD) HMMs of 5-states, left to right, no-skip topology with diagonal covariance matrix are constructed. Conventional MDL-based decision tree is applied to do model clustering[10]. The penalty scaling factor α is set to 1. The number of spectral senones, or the number of terminal leaves of the spectral decision tree is 1,755. A database of 1,000 English utterances ($\sim$ 1 hour) uttered by the source speaker is also used for training the English GMM-HMM based TTS. The number of spectrum senones is 1,818.

Wall Street Journal CSR corpus is used to train CD-DNN-HMM acoustic model. Training set (SI-284) contains 78 hours utterances recorded by 284 native American English speakers. A context dependent GMM-HMM models (CD-GMM-HMM) are first trained in the ML sense with subset of the training data (SI-84) which contains 15 hours utterances of 84 speakers. The acoustic features, extracted by a 25ms hamming window, shifted every 10ms, consist of 38 MFCCs plus log energy. Three states, left-to-right HMM triphone models, each state with 16 Gaussians components, diagonal covariance distribution, are trained. The phone set is constructed by grouping TIMIT phonemes into 40 phonemes. The total number of "senones" after state-tying is 2,754.

Acoustic models are then enhanced by DNN training with all training data (SI-284)[13]. Our DNN model (CD-DNN-HMM) is a 6 layer network, consisting of 1 input layer, 4 hidden layers, each layer with 2K units, and 1 output layer, with the same number of senones output as in CD-GMM-HMM. The input of DNN is MFCCs, which contains 5 left frames, the current frame and 5 right frames (429 dimensions). Each dimension is normalized to zero mean and unit variance. Our DNN is initialized with the Deep Belief Network (DBN) pre-training procedure [12]. All weights and bias are then discriminatively tuned using about 100 epochs in the BP phase.And the learning rates and size of mini-batch are also set for each RBM and in each training phase. The state transition parameters are obtained from original CD-GMM-HMM training.

4.2. Experimental Results and Analysis

Four systems have been built for evaluating the cross-lingual TTS performance of our proposed KLD-DNN approach. They are two KLD-DNN based systems, including: ***System*** I constructed with senone mapping and ***System*** II constructed with frame mapping and two reference systems, including: ***System*** III, our earlier "trajectory tiling" based TTS [2] where a manually set bilinear spectral warping was used to equalize the speaker difference between the source and reference speakers; ***System*** IV, baseline GMM-HMM parametric TTS trained on 1,000 recorded Mandarin sentences of the bi-lingual source speaker; this system is used as the upper bound of the CL-TTS performance since it is built directly with the recorded sentences in the target language, i.e., Mandarin, but it is not realistic in practice.

4.2.1. Objective Test

The log-spectral distortion between the synthesized L2 speech and the source speaker's L2 natural recordings are shown in Table1. Our proposed methods, both ***System*** I which is based on senone mapping and ***System*** II which is based on frame mapping, outperform the "trajectory tiling" approach significantly in terms of LSD.

Table 1. LSD(dB) on test set

	I	II	III	IV
LSD(dB)	4.68	4.50	5.39	3.91

4.2.2. Intelligibility Test

An informal intelligibility test was conducted to evaluate the the four systems. It is informal because it was not subjectively tested with semantically unmeaningful sentences (SUS). However, the sentences used in the test are not that common like everyday greeting sentences. In other words, significant effort is still needed by the subjects to transcribe the testing utterances phonetically correct. Five native Mandarin speakers with normal hearing were asked to transcribe 20 testing sentences randomly chosen from a set of 100 synthesized sentences. And the intelligibility test result is shown in Table 2. According to the table, all four systems, based upon the GMM-HMM parametric TTS framework, which is well known for its intrinsic high intelligibility, performed well in the intelligibility test. The difference between any two systems is statistically insignificant.

Table 2. Intelligibility score(%) for 20 synthesized sentences

	I	II	III	IV
Intelligibility	98.1	97.9	98.2	98.7

4.2.3. Naturalness and Speaker Similarity Subjective Test

A total of 10 native Mandarin speakers with normal hearing participated in the naturalness and speaker similarity preference test. In the naturalness test they were asked to judge 20 synthesized sentences in a five-point scale MOS[14]: 5-excellent, 4-good, 3-fair, 2-poor, 1-bad. While in the similarity test, they were asked to give subjective opinions in terms of a five-point scale DMOS and the synthesized sentences were given side-by-side with the corresponding original recordings of the source speaker. The 5-point DMOS scores are: 5-very similar, 4-quite similar, 3-similar, 2-different, 1-very different.

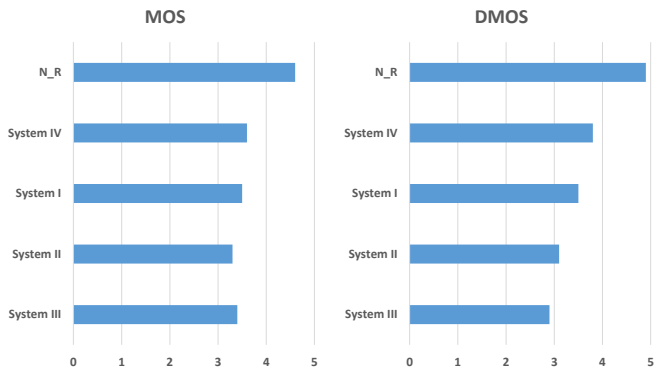

Fig. 3. Naturalness (MOS) and similarity (DMOS) scores, N_R is source speaker's Mandarin natural recordings

As indicated by the results depicted in Fig.3, ***System*** I based upon senone mapping performs much better than best ***System*** III previously proposed by us with a DMOS advantage of 0.6 in speaker similarity, or 3.5 vs 2.9. In MOS scores, ***System*** I achieves a slightly better rating than ***System*** III, i.e., 3.5 vs 3.4. MOS naturalness scores and DMOS speaker similarity scores also indicate ***System*** I is approaching the unreachable upper bound performance of the reference ***System*** IV, which is trained directly with the recorded sentences in Mandarin, while ***System*** I is trained via a cross-lingual training. The MOS difference is 3.5 vs 3.6 while the DMOS difference is 3.5 vs 3.8. This is very satisfactory since as far as we know, no CL-TTS has ever reached such a high speaker similarity performance while still keeping a very decent MOS naturalness socre and good intelligibility. Without using the transcription, ***System*** II, based upon the same KLD-DNN approach but with a statistically soft "frame mapping", achieves MOS score of 3.3 in naturalness and DMOS scores of 3.1 in speaker similarity. Given the fact that no transcriptions are required for proper segmentation of the training data into appropriate "senone" chunks plus other probability to improve the overall system performance, ***System*** II has a high potential for CL-TTS training of personalized voice. Overall, the good performance of both ***System*** I and ***System*** II really demonstrate the power of speaker independent neural net for equalizing the speaker and language differences, the two most challenging issues in CL-TTS training, and the KLD's power in "aligning" speech units phonetically down to the sub-phonemic senone or frame level. Some samples of the synthesized utterances are give on the web link: http://feng-long.github.io/CL-TTS.

5. CONCLUSIONS

In this study, we propose to use SI-DNN to equalize the speaker difference in different languages and KLD to measure the phonetic distortion between two given acoustic segments at senone or frame level probabilistically for training high performance TTS cross-lingually with good intelligibility, naturalness and high speaker similarity. The senone mapping based CL-TTS thus trained has been shown the effectiveness of the proposed approach. The MOS of naturalness has achieved 3.5 and DMOS of speaker similarity has also achieved 3.5 by the senone mapping based CL-TTS; those two scores are approaching the corresponding scores of MOS and DMOS (3.6 and 3.8) achieved by the TTS directly trained by the recorded sentences in the target language of the same source speaker.

6. REFERENCES

[1] Y. Qian, H. Liang, F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009

[2] J. He, Y. Qian, F. K. Soong, S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language TTS," in *Proceedings Interspeech*2012.

[3] H. Wang, F. K. Soong, H. Meng, "A spectral space warping approach to cross-lingual voice transformation in HMM-based TTS," in *Proceedings ICASSP*, pp.4874–4878, 2015

[4] J. Latorre, K. Iwano, S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proceedings ICASSP*, pp.1–4, 2005

[5] L. Badino, C. Barolo, S. Quazza, "Language independent phoneme mapping for foreign TTS," in *Porceedings 5th ISCA speech synthesis workshop*, pp. 271–218, 2004.

[6] G. E. Dahl, D. Yu, L. Deng, A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recgnition," in *IEEE Transactions on Audio, Speech and Language Processing*. vol. 20, no. 1, pp. 30–42,2012

[7] F. Seide, G. Li, D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proceedings Interspeech*, pp. 437–438, 2011

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings ICASSP*, pp. 751–756, 1993.

[9] F. K. Soong and B.H. Juang, "Line Spectrm Pair (LSP) and speech data compression," in *Proceddings ICASSP*, pp. 1.10.1–1.10.4, 1984.

[10] K. Shinoda, T. Watanabe, "Acoustic modeling based on the mdl principle for speech recognition," in *Proceeding EuroSpeech*, pp. 99–102, 1997.

[11] K. Liu, J. Zhang, Y. Yan, "High quality voice conversion through phoneme based linear mapping functions with STRAIGHT for Mandarin," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery*, vol. 3, pp. 410–414, 2007.

[12] G. E. Hinton, S. Osindero, Y. W. Teh, "A fast learning algorithm for deep belief nets," in *Neural Computation*, vol. 18, no. 7, pp. 1527–1544, 2006

[13] W. Hu, Y. Qian, F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning(CALL)," in *Proceedings Interspeech*, pp. 1886–1890, 2013.

[14] "Methods for Subjective Determination of Transmission Quality," Rec.P.800,ITU-T Std., 1996.

[15] S. Kullback, R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951