Objective Evaluation of HMM-based Speech Synthesis System Using Kullback-Leibler Divergence

C.-T. Do[1]*, M. Evrard*[1]*, A. Leman*[2]*, C. d'Alessandro*[1]*, A. Rilliard*[1] *and J.-L. Crebouw*[2]

[1]LIMSI-CNRS, B.P. 133, F-91403 Orsay Cedex, France
[2]Vocally Inc., 21 Rue des Hayeps, F-93100 Montreuil, France
E-mails: {ctdo, evrard, cda, rilliard}@limsi.fr, {adrien.leman, jl.crebouw}@vocally.fr

Abstract

In this paper, we propose a new objective evaluation method for hidden Markov model (HMM)-based speech synthesis using Kullback-Leibler divergence (KLD). The KLD is used to measure the difference between the probability density functions (PDFs) of the acoustic feature vectors extracted from natural training and synthetic speech data. For the evaluation, Gaussian mixture model (GMM) is used to model the distribution of acoustic feature vectors, including the fundamental frequency (F0). Continuous F0, obtained with linear interpolation, is used in the evaluation. In essence, the KLD is the expectation of the logarithmic difference between the likelihoods calculated on training and synthetic speech. This likelihood difference is appropriate to characterize the quality of a HMM-based speech synthesis system in generating synthetic speech using a maximum likelihood criterion. The objective evaluation is tested with 3 different HMM-based speech synthesis systems which use multi-space distribution (MSD) to model discontinuous F0. These systems are trained on a common speech corpus in French. We propose an index to evaluate HMM-based speech synthesis system which takes into account the relative variation of the KLDs on test sets of synthetic and natural speech. This index correlates inversely with the result of the MOS (mean opinion score) perceptual test.

Index Terms: HMM-based speech synthesis, Objective evaluation, Kullback-Leibler divergence, Gaussian mixture model, Fundamental frequency (F0)

1. Introduction

Evaluation of speech synthesis system is an active research topic in speech synthesis. Subjective evaluation is usually used as the principal quality measure of a speech synthesis system. However, performing subjective tests is time consuming and, sometimes, the result is not perfectly reproducible. In contrast, objective evaluation can be automatically done and is completely reproducible. There are a number of objective measures which can be used to evaluate the quality of a speech synthesis system, for instance the PESQ (Perceptual Evaluation of Speech Quality) [1] or the speech intelligibility prediction based on the model of auditory pre-processing [2]. These criteria are perceptually-motivated, i.e. the human speech perception mechanism is taken into account in the design of the objective evaluation. Taking into account the perception of synthetic speech makes it possible to ascertain a given correlation between objective and subjective evaluations.

Currently, the intelligibility of speech generated by speech synthesis systems, and particularly by HMM-based speech synthesis system, is rather high [3]. The focus of the evaluation of HMM-based synthetic speech has therefore shifted towards how closely a synthetic voice mirrors a human voice [3]. In speech synthesis system based on HMMs, e.g. HTS [6, 7], the spectral and excitation parameters for synthesizing synthetic speech are generated from statistical models, namely the HMMs, which are trained on natural speech. The models parameters are trained by maximizing the likelihood of the models on training data. During the parameters generation process, given a word sequence, the acoustic parameters are generated by maximizing the output likelihood given the models parameters, obtained from training. Therefore, evaluating HMM-based speech synthesis using a criterion which takes into account the likelihood of the training and synthetic data is appropriate. On the other hand, the difference of likelihoods calculated on natural training and synthetic speech might mirror the difference between natural and synthetic speech. To the best of our knowledge, there is not yet an objective evaluation for HMM-based speech synthesis system based on this criterion.

In this paper, we propose a new objective evaluation method of HMM-based speech synthesis system which takes into account the likelihood difference between natural training and synthetic speech. This objective evaluation makes use of the Kullback-Leibler divergence [8]. Kullback-Leibler divergence (KLD) is a measure of dissimilarity between two probability distributions. It has been successfully applied to outline the mismatch between training and test conditions in automatic speech recognition [4, 5]. In our proposition, the KLD will be used to characterize how well a HMM-based speech synthesis system generates synthetic speech that mirrors natural speech. The experiments are performed on HMM-based speech synthesis systems which use maximum likelihood criterion [6].

The paper is organized as follows. Section 2 describes the principle of the proposed objective evaluation method using the KLD. The speech corpus for the experiments is introduced in section 3. Section 4 presents the HMM-based speech synthesis systems. Objective evaluation results are introduced in section 5. The result of a MOS (mean opinion score) perceptual test is presented in section 6. Finally, section 7 concludes the paper and introduces perspectives of the work.

2. Objective Evaluation using the Kullback-Leibler Divergence

2.1. Kullback-Leibler divergence

In HMM-based TTS (text-to-speech), during the synthesis process, an arbitrary given text to be synthesized is first converted to a context-based labels sequence. After that, the context-dependent HMMs are concatenated to build the sentence HMMs from the labels sequences. The state durations of

the HMMs are calculated by maximizing the output likelihood of the state durations. In a similar manner, the acoustic feature vectors, including the dynamic features, are calculated by maximizing the output likelihood of the sentence HMMs using the speech parameter generation algorithm [9]. Indeed, maximum likelihood is the essential criterion which is used in state-of-the-art HMM-based TTS, including the generation of the acoustic feature vectors [6].

The KLD between two probability distributions, or two probability density functions (PDFs) $f(\mathbf{x})$ and $g(\mathbf{x})$ if the variable $\mathbf{x}$ is continuous, is defined as follows:

$$D(f(\mathbf{x}), g(\mathbf{x})) = \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \mathrm{d}\mathbf{x} \quad (1)$$

where $f(\mathbf{x})$ is the reference PDF. In essence, the KLD is the expectation of the logarithmic difference between the likelihoods calculated with $f(\mathbf{x})$ and $g(\mathbf{x})$. If the PDFs $f(\mathbf{x})$ and $g(\mathbf{x})$ are the global PDFs of the natural and synthetic speech, respectively, this divergence would be able to measure how well the synthetic speech mirrors the natural one, based on the difference of the global likelihoods.

2.2. GMM-based acoustic modeling for objective evaluation

In general, an acoustic feature vector $\mathbf{x}$ of an HMM-based TTS system consists of spectral and excitation features along with their dynamic features (delta and delta-delta coefficients) [6]. In a standard HMM-based TTS system, the modeling of the fundamental frequency (F0) is not trivial due to the discontinuity of F0 values across unvoiced regions. Generally, the F0 values in the unvoiced regions cannot be estimated, using standard F0 estimation algorithms [10, 11], and they are assumed to be undefined in these regions. The discontinuity of F0 makes it impossible to correctly model it using a simple continuous distribution, for instance the Gaussian Mixture Model (GMM).

The multi-space distribution HMM (MSD-HMM) [12] provides a solution to this problem by using a combination of discrete and continuous distributions and it is now the default modeling approach in state-of-the-art HMM TTS system [6, 7]. Within the MSD-HMM modeling framework, the state output distribution $b(f_0)$ of the F0 has the following form [13]:

$$b(f_0) = \begin{cases} \lambda \mathcal{N}(f_0, \mu, \sigma) & \text{if } f_0 \in \text{voiced region} \\ 1 - \lambda & \text{if } f_0 \in \text{unvoiced region} \end{cases}$$

where f_0 is the observation of F0, λ and $1-\lambda$ are the probabilities of voiced and unvoiced regions, and μ and σ are the means and variances of the Gaussian distribution of F0 in the voiced regions. Good speech synthesis performance could be achieved with the MSD-HMM. However, if the distribution of the acoustic feature vectors is modeled with MSD, the calculation of the KLD would be not readily feasible since, to our knowledge, there are no existing algorithms which permit to calculate the KLD between the MSDs.

In this paper, for the sake of the calculation of the KLD, we simplify the acoustic modeling within the objective evaluation by using the Gaussian mixture model (GMM) to model the distribution of the acoustic feature vectors $\mathbf{x}$. Hence, the form of $f(\mathbf{x})$ is defined as:

$$f(\mathbf{x}) = \sum_{\delta=1}^{M} \lambda_\delta \mathcal{N}(\mathbf{x}, \mu_\delta, \mathbf{\Sigma}_\delta) \quad (2)$$

for $\mathbf{x} \in$ voiced and unvoiced regions. In equation (2), $\lambda_\delta, \mu_\delta$ and Σ_δ are the prior probability, mean vector and covariance matrix, respectively, of the δ^{th} multivariate Gaussian component $\mathcal{N}(\mathbf{x}, \mu_\delta, \mathbf{\Sigma}_\delta)$ in the GMM consisting of M Gaussian components. Using GMM to model the distribution of the acoustic feature vectors needs the F0 (and the Δ and $\Delta\Delta$ of F0) to be continuous. To this end, we use linear interpolation to interpolate the values of F0, and the Δ and $\Delta\Delta$ of F0, in the unvoiced regions. A continuous F0 has advantages. Indeed, it has been shown that an HMM based speech synthesis system using a continuous F0 produces more expressive F0 contours than one based on the MSD [13, 14]. However, in this work, continuous F0 and GMMs are used only for the evaluation. Inside the HMM-based TTS systems, the MSDs are used for acoustic modeling. An example of continuous F0 values of an utterance, obtained after linear interpolation, is shown in Fig. 1. The original F0 values (blue line) is taken from the acoustic feature vectors extracted by an HMM-based TTS system. The continuous F0 values (red line) are obtained by linear interpolating the original F0 values. The continuous F0 is modeled with GMMs for the objective evaluation using KLD.

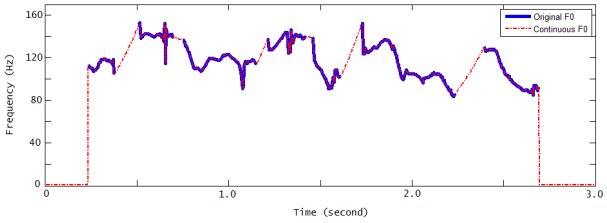

Figure 1: Example of continuous F0 (red line), obtained by linearly interpolating the discontinuous one (blue line), from an utterance in the speech corpus. Continuous F0 is used only for the objective evaluation.

2.3. Kullback-Leibler divergence between GMMs

The calculation of the KLD between the GMMs is not analytically tractable. Amongst the currently available methods [15], Monte Carlo sampling is the sole method that can calculate the KLD between GMMs with arbitrary accuracy. We thus apply the Monte Carlo sampling method to calculate the KLDs between the acoustic feature vectors global PDFs which are modeled with GMMs. Indeed, the equation (1) can be rewritten as

$$D(f(\mathbf{x}), g(\mathbf{x})) = \mathbb{E}_f \left[\log \frac{f(\mathbf{x})}{g(\mathbf{x})} \right]$$

where $\mathbb{E}$ is the mathematical expectation. This expectation can be approximated by generating N independent identically distributed (i.i.d) random vectors $\mathbf{x}_i, i = 1, \ldots, N$ following the PDF $f(\mathbf{x})$, and then, by calculating the empirical mean of $\log[f(\mathbf{x}_i)/g(\mathbf{x}_i)], i = 1, \ldots, N$ since

$$\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{f(\mathbf{x}_i)}{g(\mathbf{x}_i)} \right] \rightarrow \mathbb{E}_f \left[\log \frac{f(\mathbf{x})}{g(\mathbf{x})} \right]$$

when $N \rightarrow +\infty$. To generate a random i.i.d vector following the PDF $f(\mathbf{x})$ which is a GMM, a random indicator $\delta \in 1, \ldots, M$ is generated following the a priori probabilities $\lambda_k, k = 1, \ldots, M$. Then, the random vector $\mathbf{x}_i$ is generated from the corresponding Gaussian component $\mathcal{N}(\mathbf{x}, \mu_\delta, \mathbf{\Sigma}_\delta)$.

3. Speech Corpus

The experiments in this paper use a natural speech corpus in French consisting of 1155 utterances. The sentences in the corpus are phonetically balanced according to phones, diphones

and triphones segmentations. All the sentences in the corpus have been recorded with a native male voice in a slow reading, neutral expression with moderate hyper-articulation. The utterances have been recorded using a dynamic microphone with a cardioid polar pattern at sampling frequency of 48 kHz and 16 bits resolution. Subsequently, the utterances have been downsampled to 16 kHz. The corpus has been labeled either automatically by using the E-HMM (ergodic HMM) labeler of Festvox [17] or manually by a human. The average duration of each utterance in the corpus is around 3 seconds. The corpus is divided into training and test sets. The training set consists of 1115 utterances and the test set consists of 40 utterances. These two sets are not overlapped.

4. HMM-based TTS Systems

The KLD is applied to evaluate the quality of three different French HMM-based TTS systems trained on this corpus. These 3 HMM-based TTS systems have been implemented based on the HTS (HMM-based speech synthesis system) toolkit [7]. They are denoted as TTS_1, TTS_2 and TTS_3. The 3 systems share the same standard structure, including the natural language processing (NLP), training and synthesis components [16]. In this work, the systems TTS_1 and TTS_3 are implemented within the Festival TTS platform [18] using the labels aligned manually. The only difference between the implementation of these two systems is the amount of acoustic training data. The system TTS_1 uses 1115 utterances in the training corpus for training acoustic models. The system TTS_3 uses only 300 out of the 1115 utterances in the training corpus for acoustic models training. Indeed, the purpose is to evaluate the impact of the amount of acoustic training data on the quality of the HMM-based TTS systems. The TTS_1 and TTS_3 systems use a common NLP component which is adapted from the Festival TTS platform to French language. This NLP component extracts 49 contextual factors which are inspired from [19] and adapted to French language. In fact, from 53 general contextual factors proposed in [19], the 7 factors related to the stress lexical accents have not been used. On the other hand, we add 3 contextual factors related to the syllable situation, for instance: end of sentence, breathing, end of word, end of word comprising a liaison with the next word, middle of word. There are finally 49 contextual factors in total.

The system TTS_2 is trained with 1115 utterances from the training speech corpus, the same as those used for training the system TTS_1, with the labels aligned automatically by using the E-HMM labeler. However, it differs from the TTS_1 and TTS_3 in the NLP component. Actually, the NLP component of the TTS_2 system extracts only 19, amongst 53 [19], contextual factors from the text. These 19 contextual factors were selected based on the particularities of the French language. The purpose is to evaluate the impact of the contextual factors on the quality of HMM-based TTS system [20].

The three TTS systems use the HTS toolkit [7] for training context-dependent HMMs. The training of the context-dependent HMMs utilizes the acoustic features, extracted every 5ms from the training utterances using the SPTK toolkit [21], and the labels from the training corpus. The acoustic features include mel-generalized cepstrum (MGC) coefficients, the log of F0 and their Δ and $\Delta\Delta$ coefficients. The duration is also modeled within the HMMs. In the synthesis, speech waveform is synthesized directly from the generated MGC coefficients and F0 values by using the MLSA filter [7]. The speech synthesis component of each TTS system takes the contextual factors extracted from input text using the corresponding NLP.

5. Objective Evaluation Results

5.1. Objective evaluation experimental protocol

Each TTS system is used to synthesize 40 synthetic utterances having the same linguistic contents of the utterances in the test set consisting of 40 natural utterances (see section 3). These 40 utterances have the same nature of those used for the training, i.e. the utterances are phonetically balanced and have 3-second average duration. The total duration of 40 testing utterances is around 2 minutes. In the evaluation, for each TTS system, acoustic feature vectors are extracted from the training and synthetic data every 5ms. An acoustic feature vector $\mathbf{x}$ has 108 dimensions which consist of 35 MGC coefficients, the $\log(F0)$ along with their Δ and $\Delta\Delta$ coefficients. From the training corpus of 1115 utterances, around 711K acoustic feature vectors are extracted for estimating the reference PDF $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ of the system TTS_1 and TTS_2, respectively. With the system TTS_3 using 300 utterances for training, around 185K feature vectors have been extracted from the training speech to estimate its reference PDF $f_3(\mathbf{x})$.

From the 40 synthetic utterances, around 24K acoustic feature vectors are extracted. Around 24K feature vectors have also been extracted from the 40 natural utterances. The feature vectors extracted from the sets of 40 synthetic utterances, synthesized by the systems TTS_1, TTS_2 and TTS_3, are used to estimate the PDFs $g_{S1}(\mathbf{x})$, $g_{S2}(\mathbf{x})$ and $g_{S3}(\mathbf{x})$, respectively. The feature vectors extracted from the natural speech set, using the feature extraction of the system TTS_1, TTS_2 and TTS_3, are used to estimate the PDFs $g_{N1}(\mathbf{x})$, $g_{N2}(\mathbf{x})$ and $g_{N3}(\mathbf{x})$, respectively. The acoustic feature vectors, extracted during the evaluation, are the same as those extracted and generated during the training and synthesis, i.e. the F0 is undefined in unvoiced regions. However, the feature vectors, extracted from the same natural speech set by using different feature extractors, might be slightly different since the TTS systems are developed at different sites.

The PDFs $f(\mathbf{x})$ and $g(\mathbf{x})$ are modeled by using GMMs of 16 Gaussian components with diagonal covariance matrices ($M = 16$). As mentioned in section 2.2, the F0 is linearly interpolated to be continuous so that the acoustic modeling using the GMMs is applicable for the evaluation. For $i = \{1, 2, 3\}$, the KLDs D_{Si} are calculated between $f_i(\mathbf{x})$ and $g_{Si}(\mathbf{x})$, and the KLDs D_{Ni} are calculated between $f_i(\mathbf{x})$ and $g_{Ni}(\mathbf{x})$. The KLD between the GMMs are calculated by using the Monte Carlo sampling method (see section 2.3). In our experiments, $N = 100K$ vectors are randomly generated following the reference PDF $f_i(\mathbf{x})$ of each TTS system and the KLDs are calculated by using these vectors. The expectation-maximization (EM) algorithm [22] is applied once on the training data to estimate the GMMs $f_i(\mathbf{x}), i = 1, 2, 3$. However, the EM algorithm is initialization-dependent [22]. Therefore, in each KLD calculation, e.g. the D_{Si} between $f_i(\mathbf{x})$ and $g_{Si}(\mathbf{x})$, we apply the EM algorithm 50 times on the same test data (synthetic data of the i^{th} TTS system) to get 50 variational versions of $g_{Si}(\mathbf{x})$. After that, the KLD D_{Si} is defined as the average of the 50 KLDs calculated between $f_i(\mathbf{x})$ and the 50 variational versions of $g_{Si}(\mathbf{x})$. The standard deviation of D_{Si} is also calculated from these 50 KLDs.

5.2. Numerical results

Figure 2 shows the numerical results of the KLDs calculated for three systems, TTS_1, TTS_2 and TTS_3. It can be observed that, in each TTS system, the KLD of synthetic speech is larger than that of the natural speech. That is, the KLD, calculated

in this manner, has the ability to distinguish between synthetic and natural speech. Further, it can be observed that the D_{N1} and D_{N2} are different even though the systems TTS_1 and TTS_2 use the same sets of natural speech for training and test, and the same type of acoustic feature vector is used. This difference can be explained from the fact that the TTS systems have been developed in different sites with slight differences in the pre-processing of the audio data. However, these differences raise the difficulties in comparing directly the KLDs calculated on different TTS systems.

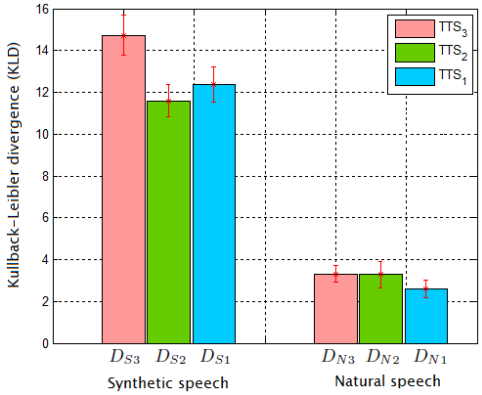

Figure 2: Kullback-Leibler divergence (KLD) calculated between the PDFs of the training and test (synthetic and natural) speech of three systems TTS_1, TTS_2 and TTS_3. Error bars represent the standard deviations.

5.3. System evaluation index

We thus propose a system evaluation index (SEI) I, based on the KLDs calculated on synthetic and natural speech of an HMM-based TTS system, to evaluate the overall quality of the system. The expression of I is as follows:

$$I = D_N \left(1 - \frac{D_N}{D_S}\right) \tag{3}$$

where D_N and D_S are the KLD calculated between the reference PDF $f(\mathbf{x})$ and the PDFs $g_N(\mathbf{x})$ and $g_S(\mathbf{x})$ of test sets consisting of natural and synthetic speech, respectively. In equation (3), D_N is used as a multiplication factor since it represents the quality of the acoustic training data and of the acoustic feature vectors used by a TTS system. The smaller the D_N, the better the TTS system. On the other hand, the ratio $\frac{D_N}{D_S} \leq 1$ since $D_N \leq D_S$ in general (and also as observed empirically). The closer the D_S moves towards the D_N, the better the TTS system. Therefore, the smaller the index I, the better the TTS system. In an ideal system, $I = 0$ when $D_S = D_N$.

Calculating the index I_i using the values D_{Ni} and D_{Si}, $i = \{1, 2, 3\}$ of the three systems TTS_1, TTS_2 and TTS_3, we have $I_1 = 2.05$, $I_2 = 2.35$ and $I_3 = 2.57$, i.e., the order of the three systems following their overall quality is: $TTS_1 > TTS_2 > TTS_3$.

6. MOS Perceptual Test

A MOS (mean opinion score) perceptual test is performed to evaluate subjectively the quality of the three TTS systems. From 40 utterances of synthetic speech synthesized by each TTS system, 21 utterances are selected for the perceptual test. The selected utterances are the same for three TTS systems and natural speech. Amongst 21 utterances, one is used at the beginning of the listening session in order to present the listeners with the quality of the sounds they will have to judge. The 20 remaining utterances are used for the perceptual test. From three TTS systems and the natural speech, 84 utterances were selected and presented to listeners in a random order (the scores of the first 4 utterances for familiarization were not counted).

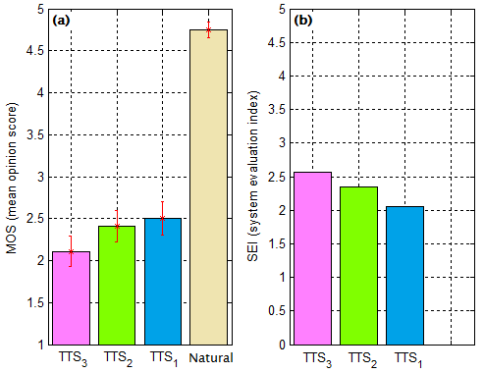

Figure 3: MOS scores (a) and the system evaluation indices I (b) of three systems TTS_1, TTS_2 and TTS_3. Error bars of the MOS scores indicate 95% of the confidence intervals.

Nine subjects participated in the test. They had to rate the overall quality of each utterance, based on a scale from 1 to 5 (1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent). A two-way ANOVA was run on the result. The two factors were the TTS system (4 levels) and the utterances (20 levels). The two factors have significant effects, as single factor, on the MOS score. The interaction is not significant. The TTS system factor explains most of the observed variance ($\eta^2 = 0.63$). A post-hoc Tukey test groups the systems into 3 groups of homogeneous levels of perceived quality, ordered by descending levels of the MOS scores: natural speech, TTS_1 and TTS_2, and TTS_3. The MOS scores and the system evaluation indices (SEIs) I are shown in Fig. 3. It can be observed from the Fig. 3 that, the SEIs correlates inversely with the MOS scores (correlation coefficient equals -0.93). Indeed, the smaller the SEI, the better the system, whereas, the larger the MOS score, the better the system.

7. Conclusion and Perspectives

We have proposed to use Kullback-Leibler divergence for the evaluation of the overall quality of HMM-based speech synthesis system. The KLD is calculated between the global PDFs, which are GMMs, of the acoustic feature vectors extracted from natural training and synthetic speech data. Experimental results have shown that the KLD calculated on natural speech is substantially smaller compared to that calculated on synthetic speech synthesized by an HMM-based TTS system. We have proposed a system evaluation index (SEI) to evaluate the overall quality of an HMM-based TTS system. This index is based on the KLD, calculated on a test set of natural speech, and the relative variation of the KLDs calculated on test sets of synthetic and natural speech. Experimental results have shown that the SEI correlates inversely with the MOS score obtained by perceptual test. Future work would investigate the use of this method to evaluate HMM-based TTS systems using different vocoders, for instance STRAIGHT [23].

Acknowledgments: This work has been co-financed by OSEO, the French State Agency for Innovation, and the Région Île-de-France under the FUI project ADN T-R on the fabrication of digital double.

8. References

[1] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", *ITU-T Recommendation*, February 2001.

[2] Christiansen, C., Pedersen, M. S. and Dau, T., "Prediction of speech intelligibility based on an auditory preprocessing model", *Speech Communication*, vol. 52, no. 7-8, pp. 678-692, August 2010.

[3] Campbell, N., "Evaluation of speech synthesis: from reading machines to talking machines", *Evaluation of Text and Speech Systems* (L. Dybkjoer et al. Eds.), Chap. 2, pp. 29-64, 2007.

[4] Do, C.-T., Pastor, D., and Goalic, A., "A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech", *Speech Communication*, vol. 54, no. 1, pp. 119-133, Jan. 2012.

[5] Do, C.-T., Taghizadeh, M.J., and Garner, P.N., "Combining cepstral normalization and cochlear implant-like speech processing for microphone array-based speech recognition", *2012 IEEE Workshop on Spoken Language Technology (SLT)*, pp. 137-142, Miami, USA, 2-5 Dec. 2012.

[6] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K., "Speech synthesis based on hidden Markov models", *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[7] Tokuda, K., Zen, H., and Black, A., "An HMM-based speech synthesis system applied to English", *IEEE Workshop on Speech Synthesis*, September 2002.

[8] Kullback, S. and Leibler, R. A., "On information and sufficiency", *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79-86, March 1951.

[9] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", *IEEE ICASSP'00*, vol. 3, pp. 1315-1318, June 2000.

[10] Boersma, P., "Praat, a system for doing phonetics by computer", *Glot International*, vol. 5:9/10, pp. 341-345, 2001.

[11] de Cheveigné, A., and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917-1930, April 2004.

[12] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., "Multi-space probability distribution HMM", *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455-464, March 2002.

[13] Yu, K., Toda, T., Gasic, M., Keizer, S., Mairesse, F., Thomson, B., and Young, S., "Probabilistic modelling of F0 in unvoiced regions in HMM based speech synthesis", *IEEE ICASSP'09*, pp. 3773-3776, Taipei, Taiwan, April 19-24, 2000.

[14] Yu, K., and Young, S., "Continuous F0 modelling for HMM based statistical parametric speech synthesis", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1071-1079, 2011.

[15] Hershey, J.R., and Olsen, P.A., "Approximating the Kullback-Leibler divergence between Gaussian mixture models", *IEEE ICASSP'07*, vol. 4, pp. 317-320, April 2007.

[16] Nguyen, T.T.T., d'Alessandro, C., Rilliard, A., and Tran, D.-D., "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation", *INTERSPEECH'13*, pp. 2311-2315, Lyon, France, August 25-29, 2013.

[17] Black, A., and Lenzo, K.A., "Building synthetic voices", http://festvox.org/bsv, 2007.

[18] Taylor, P., Black, A.W., and Caley, R., "The architecture of the Festival speech synthesis system", *3rd ESCA/COCOSDA Workshop on Speech Synthesis*, NSW, Australia, November 26-29, 1998.

[19] HTS Working Group, "An example of context-dependent label format for HMM-based speech synthesis in English", *The HTS CMU-ARCTIC Demo*, 2012.

[20] Le Maguer, S., Barbot, N., and Boeffard, O., "Evaluation of contextual descriptors for HMM-based speech synthesis in French", *8th ISCA Speech Synthesis Workshop*, pp. 153-158, Barcelona, Spain, September 2013.

[21] Kobayashi, T., Tokuda, K., Koishida, K., et al., "Speech signal processing toolkit (SPTK)", *Version 3.7*, December 25th 2013.

[22] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[23] Kawahara, H., "STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds", *Acoust. Sci & Tech.*, vol. 27, no. 6, pp. 349-353, 2006.