# Measuring Attribute Dissimilarity with HMM KL-Divergence for Speech Synthesis

*Yong ZHAO[1], Chengsuo ZHANG[2], Frank K. SOONG[1], Min CHU[1] and Xi XIAO[2]*

[1] Speech Group, Microsoft Research Asia, China
[2] Department of Electronic Engineering, Tsinghua University, China
{yzhao, frankkps, mchu}@microsoft.com

## Abstract

This paper proposes to use KLD between context-dependent HMMs as target cost in unit selection TTS systems. We train context-dependent HMMs to characterize the contextual attributes of units, and calculate Kullback-Leibler Divergence (KLD) between the corresponding models. We demonstrate that the KLD measure provides a statistically meaningful way to analyze the underlining relations among elements of attributes. With the aid of multidimensional scaling, a set of attributes, including phonetic, prosodic and numerical contexts, are examined by graphically representing elements of the attribute as points on a low dimensional space, where the distances among points agree with the KLDs among the elements. The KLD between multi-space probability distribution HMMs is derived. A perceptual experiment shows that the TTT system defined with the KLD-based target cost sounds slightly better than one with the manually-tuned.

**Index Terms:** speech synthesis, unit selection, target cost, Kullback-Leibler divergence, HMM, multi-space probability distribution, multidimensional scaling

## 1. Introduction

Text-to-Speech (TTS) systems based on unit selection feature advantages in synthesizing highly natural and intelligible speech, and have become dominant in commercial applications. These systems rely on a very large database of segmental samples, where the best segment sequence is retrieved for generating speech output with the criterion to minimize a cost function. The cost function is a summation of two sub-cost functions: a concatenation cost, which reflects how well two segments concatenate, and a target cost, as is our interest in this paper, which describes the difference between target and candidate segments.

In the literature, various techniques have been proposed to define the target cost function. A number of approaches presented to minimize the generation error of synthesized speech [1][2], where costs are tuned toward minimizing the distortion of synthetic utterances from their natural counterparts as a reference. Other approaches were based on agreement with human perception [3][4][5], where synthesized utterances are scored subjectively, and costs producing a maximum correlation with subjective scores are regarded as objectively optimal.

Anyhow, in the above approaches, cost functions are optimized by means of synthesizing speech and comparing with sort of criteria. Though these approaches typically lead to a high performance in synthesis by considering all factors, including the process of the synthesis, we lack the ability to reveal the intrinsic proprieties of the target cost.

It is essential that the target cost reflect the difference between units just as human perceives [6]. In this paper, we exploit Kullback-Leibler Divergence (KLD) to estimate the target cost, where we train context-dependent Hidden Markov Models (HMM) to characterize the contextual attributes of units, and calculate KLD between these corresponding models as the distance between units.

One main advantage of the KLD measure is that it allows analyzing the underlining relations among elements of an attribute. It offers a statistically sound way to study the acoustic characteristics of the attributes from varied categories. These categories may involve phonetic, prosodic, linguistic and even paralinguistic.

In this paper we attempt to gain insight of the relations among elements of attributes with the aid of Multidimensional Scaling (MDS). A set of attributes, including phonetic, prosodic and numerical contexts, are examined by graphically representing elements of the attribute as points on a plane or line, where the distances among points agree with the KLDs among elements.

The KLD for a variety of statistical models is presented, including Multi-Space Probability Distribution (MSD) HMM. A subjective evaluation showed the system with KLD-based target cost sounds slightly better than one with manually-tuned.

This paper is organized as follows: Section 2 introduces the concepts of the KLD and its expressions for several statistical models. Section 3 describes how to exploit the KLD as a distance measure of attributes, how to evaluate its effectiveness, and its application as a target cost in unit selection systems. Experiments and discussions are given in Sectioin4 and 5 respectively.

## 2. Kullback-Leibler Divergence

The KLD between two N-dimensional probability distributions $M$ and $\tilde{M}$ [7] is defined as:

$$D(M \| \tilde{M}) = \int_{R^N} p(X|M) \log \frac{p(X|M)}{p(X|\tilde{M})} dx \quad (1)$$

KLD describes how far a "true" model $M$ is from an arbitrary model $\tilde{M}$. Note that KLD is asymmetric. If we are not sure which model is correct, we can sum up the integrals in both directions to obtain a symmetrical version of KLD:

$$D_s(M \| \tilde{M}) = D(M \| \tilde{M}) + D(\tilde{M} \| M) \quad (2)$$

When $M$ and $\tilde{M}$ are Gaussian distribution, $M \sim N(\mu, \Sigma)$ and $\tilde{M} \sim N(\tilde{\mu}, \tilde{\Sigma})$, a closed form KLD is:

$$D(M \| \tilde{M}) = \frac{1}{2}[(\mu - \tilde{\mu})^T \tilde{\Sigma}^{-1}(\mu - \tilde{\mu}) + tr(\Sigma\tilde{\Sigma}^{-1}) - \log|\Sigma\tilde{\Sigma}^{-1}| - N] \quad (3)$$

### 2.1. KLD between HMMs

HMMs are statistical models widely used in speech recognition. In [8], we derived an algorithm to assess the KLD between two general left-to-right HMMs. Given two HMMs $\mathrm{H}$ and $\tilde{\mathrm{H}}$ with parameter sets of $\{\pi, A, B\}$ and $\{\tilde{\pi}, \tilde{A}, \tilde{B}\}$ respectively, we approximate the upper bound of a symmetric KLD between two equal-length left-to-right HMMs:

$$D_s\left(\mathrm{H}\middle\|\tilde{\mathrm{H}}\right) \le \sum_{i=1}^{J-1}\left\{ l_i\left[ D\left(b_i\middle\|\tilde{b}_i\right) + \log\left(a_{ii}/\tilde{a}_{ii}\right)\right] + \tilde{l}_i\left[ D\left(\tilde{b}_i\middle\|b_i\right) + \log\left(\tilde{a}_{ii}/a_{ii}\right)\right]\right\} \quad (4)$$

Where $D(b_i \| \tilde{b}_i)$ is the KLD between the observation distributions at state *i*, $\log\left(a_{ii}/\tilde{a}_{ii}\right)$ is the log-likelihood ratio of the transition probability, and $l_i = 1/(1-a_{ii})$ is the expected duration of the $i^{\text{th}}$ state in $\mathrm{H}$.

The meaning which equation (4) suggests conforms to our intuition with it: Sum up KLDs over all states and meanwhile, the KLD between states take into account the difference of state transition and observation probabilities, both of which are weighted by the expected state duration.

For KLD between two GMMs $D(b_i \| \tilde{b}_i)$, we use unscented transform to approximate it [9].

### 2.2. KLD between multi-space probability distributions

Multi-Space Probability Distribution (MSD) was proposed by Tokuda et.al. [10][11]. The HMMs based on MSD are especially useful to model the characteristics of fundamental frequency (F0) in speech, where the voiced part is modeled in a continuous space, and the unvoiced part, a discrete symbol, is looked upon as from zero-dimensional space.

MSD assumes that the observation space $\Omega$ is composed of $G$ sub-spaces. Each sub-spaces $\Omega_g$ is of $n_g$ dimension and has a prior probability $w_g$, satisfying $\sum_{g=1}^{G} w_g = 1$. The observation is represented by a random vector $o$, which consists of two parts, the set of sub-space indices $S(o)$ and a n-dimensional random variable $V(o)$ that is distributed in all sub-spaces specified by $S(o)$. The observation probability of $o$ is defined as:

$$b(o) = \sum_{g \in S(o)} w_g N_g(V(o)) \quad (5)$$

Where $N_g(V(o))$ denotes the probability density of observation $V(o)$ for the $g^{th}$ sub-space.

Consider two MSDs consist of the same sub-spaces and have one-to-one correspondence between sub-spaces, $n_g = \tilde{n}_g$. If all $S(o)$ specify one sub-space, i.e. $|S(o)| \equiv 1$, by calculating KLD in each individual space, we get KLD between two MSDs:

$$D(b \| \tilde{b}) = D(\mathbf{w} \| \tilde{\mathbf{w}}) + \sum_{g=1}^{G} w_g D(N_g \| \tilde{N}_g) \quad (6)$$

where $D(\mathbf{w} \| \tilde{\mathbf{w}}) = \sum_{g=1}^{G} w_g \log(w_g / \tilde{w}_g)$ denotes KLD between two mixture weight vectors.

If $|S(o)| \ge 1$, i.e. some sub-spaces share their space, they literally form GMMs. We could merge these components into a super-component, and solve it by KLD between GMMs.

To estimate the KLD between MSD-HMMs, we need to substitute equation (6) for $D(b_i \| \tilde{b}_i)$ in equation (4).

## 3. KLD of Attributes

The characteristics of speech sounds are influenced by not only phonemes in place, but also contextual attributes associated with the sounds. These contextual attributes range from phonetic, prosodic, linguistic, to paralinguistic. In this paper, our research focuses on how to approximate a dissimilarity function of the speech-related attributes. The key concept is that the distances should reflect the differences of the attribute elements in respect of acoustic space. Here we propose to train context-dependent HMMs to characterize the attribute elements, and calculate KLDs between the corresponding models as the dissimilarity function of the attribute.

The first step is to train context-dependent HMMs [12]. By modifying monophones with their contextual attribute of interest, monophone transcriptions are converted to context-dependent phone transcriptions. A set of context-dependent phone models are created by copying monophones and re-estimating. Then, a decision tree based context clustering is applied to tie similar states of context-dependent phones for robust parameter estimation.

Given context-dependent HMMs, KLD is calculated between the models sharing the same central phones as dissimilarities between elements of the attribute. . Such a dissimilarity function is phone-dependent. Besides, we can calculate phone-independent dissimilarity function by averaging phone-dependent functions over all central phones.

For example, given an attribute of interest, monophones are rewritten into context-dependent phones in the form of *c:x*, where *x* is the attribute value of phone *c*. KLD between models $c{:}x_1$ and $c{:}x_2$ represents the dissimilarity between attribute element $x_1$ and $x_2$ with respect to phone *c*. The dissimilarity between $x_1$ and $x_2$ is an average of KLDs over all central phones:

$$D(x_1, x_2) = \frac{1}{N}\sum_{c \in P} D(M(c{:}x_1) \| M(c{:}x_2)) \quad (7)$$

Where $N$ is the size of phoneme set $P$.

### 3.1. Graphical interpretation of attribute KLD with multidimensional scaling

While we have in hand a KLD function for attributes, we face the problem how to evaluate the approximation goodness of the KLD measure. [13] evaluated the accuracy of KLD in terms of correlating with the divergences estimated with Monte Carlo simulation. The other approaches [9] examined by means of the performance of applications which employ KLD as the distortion measure in comparison with one without KLD.

In the paper, we adopt multidimensional scaling (MDS) [14] to graphically detect meaningfulness of KLD as dissimilarity measure. MDS is a data analysis technique that represents distances among objects as distances between points of a low-dimensional space, i.e. each object in the domain is represented by a point in the space. The points are arranged in the space so that the distances between pairs of points best approximate the distances between pairs of objects.

MDS helps reveal the underlining relations among objects. This is why we employ MDS to analyze KLD matrix for attributes. Given a KLD matrix of an attribute, elements of the attribute are projected onto a space by MDS. Assuming that KLD is a meaningful measure, elements which are close together in the space should be similar in acoustic characteristics, and elements which are far apart should be dissimilar likewise. On the other hand, if we observe that the relative locations of elements in the space agree with our knowledge with the attribute, we have reasons to believe the effectiveness of the KLD measure.

### 3.2. KLD as target cost in unit selection

One application of the proposed measure is in unit selection systems. We exploit KLD between context-dependent HMMs as the target sub-cost between target and candidate units. Let $t_i$ and $u_i$ denote the target and candidate unit. The target cost $C^t(t_i, u_i)$ is presented in the form of the sum of the KLDs between context-dependent models:

$$C^t(t_i, u_i) = \sum_{j=1}^{J} w_j^t D(M_j(t_{ij}) \| M_j(u_{ij})) \quad (8)$$

Where $M_j(t_{ij})$ denotes the model specified by unit $t_i$ in terms of its $j^{th}$ attribute, and $w_j^t$ is the weight of the $j^{th}$ sub-cost.

Note that we assume the target cost is composed of categorical attributes, such as prosodic and prosodic contexts. It holds true in a number of systems [15][16]. When a target cost involves continuous attributes, the KLD measures still work on discrete parts.

Attributes may be in form of compound. That is we take into account the interaction of multiple attributes in the target cost. One advantage of compound attributes is that the efforts to tune weights of the sub-costs $w_j^t$ are reduced. In an extreme situation, we could calculate KLD between HMMs in the context of all attributes as the target cost.

$$C^t(t_i, u_i) = D(M(t_i) \| M(u_i)) \quad (9)$$

Where $M(t_i)$ denotes the context-dependent model of target unit $t_i$.

## 4. Experiments and Results

### 4.1. Experimental setup

The Microsoft Mulan English speech corpus is used to evaluate the goodness of KLD to approximate acoustic distances for various attributes. The corpus consisted of about 6000 phonetically-balanced sentences recorded by a female voice talent. We manually annotated prosody labels on utterances, such as break levels, stress, and emphasis.

In stage of HMM training, we adopted a topology of 5-state left-to-right HMMs. Features include spectrum and F0 parameters. Spectrum features consist of 39 dimensional feature vectors (13 MFCCs, plus their delta and acceleration coefficients). F0 features consist of log F0, its delta and acceleration coefficients. Similarly, the state distribution consists of two parts: the first part models spectrum features by a single Gaussian distribution with diagonal covariance matrix; the second part models F0 features by an MSD[11]. The MSD is composed of a single Gaussian distribution with diagonal covariance matrix for voiced space and a discrete distribution outputting only one symbol, being unvoiced.

### 4.2. Contextual attributes

In this paper, the following contextual attributes are taken into account:

- **LPhC**: Left phonetic context. It consists of 40 phonemes. The phoneme set refers to one defined by Microsoft Speech SDK for American English [17].
- **RPhC**: Right phonetic context.
- **PinP**: Position of word in phrase. It takes 9 values. Values are decided by break indices surrounding the word, in the form of *n-m*, where *n* is the break index proceeding the word and *m* is the break index following. Values for the break index are chosen from the following set [18]:
  1. Word boundary.
  2. Short phrase boundary.
  3. Intonation phrase boundary.
- **PinW**: Position of syllable in word. It takes 4 values, head of word (H), middle of word (M), tail of word (T), and monosyllable (S).
- **PinS**: Position of phone in syllable.
- **Strs**: Word stress.
- **Emph**: emphasis in phrase.
- **Phns**: Number of phones in syllable. It ranges from 1 to 5. In case of more than 5 phones, set 5.
- **Syls**: Number of syllables in word. It ranges from 1 to 5. In case of more than 5 syllables, set 5.

### 4.3. Evaluation for phonetic contexts

The first experiment studies the capabilities of the KLD in capturing similarities between phonetic contexts, left phonetic context (LPhC) and right phonetic context (RPhC). Figure 1 and 2 display the planes which are transformed into by MDS from KLD matrices for LPhC and RPhC, respectively. In both graphs, we observe that phonemes, except /h/, are roughly grouped into 3 clusters:

1. Vowel.
2. Sonorant consonant. It consists of semivowels, liquids and nasals.
3. Obstruent. It consists of affricates, fricatives and stops.

Cluster Obstruent can be further subdivided into voiced and unvoiced sounds. Voiced obstruents come closer towards sonorants than does unvoiced.

In the graphs, /h/ stands apart from other phonemes. We credit it to that though phoneme /h/ in English is categorized as voiceless glottal fricative in International Phonetic Alphabet, sometimes it behaves more like a voiceless vowel due to the influence of surrounding vowels.

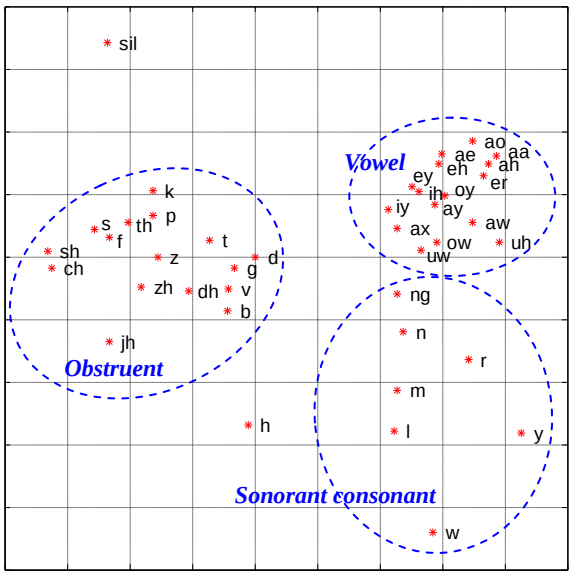


Figure 1. *MDS graph of the KLD matrix for attribute LPhC.*

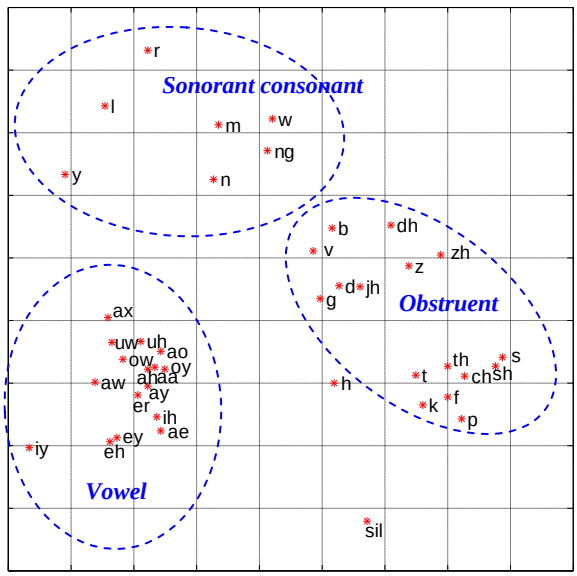


Figure 2. *MDS graph of the KLD matrix for attribute RPhC.*

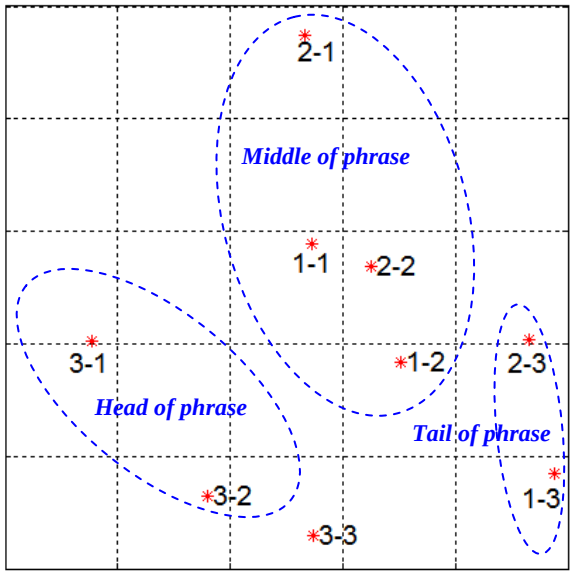


Figure 3. *MDS graph of the KLD matrix for attribute PinP.*

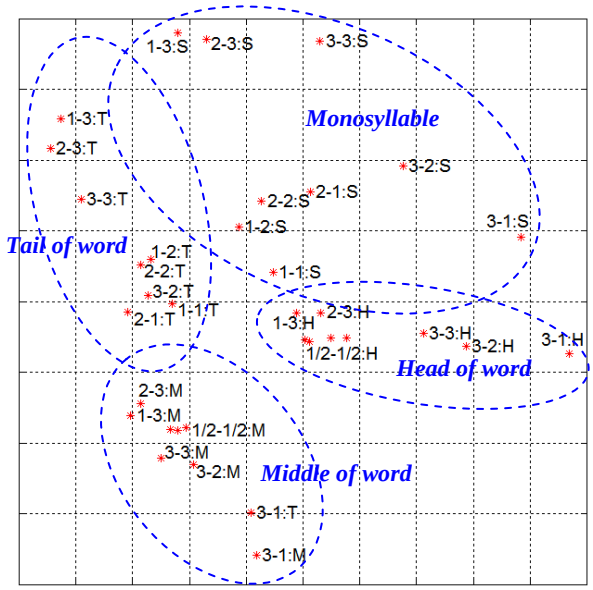


Figure 4. *MDS graph of the KLD matrix for attribute PinP*PinW.*

### 4.4. Evaluation for prosodic contexts

In this section, we examined the characteristics of KLD with respect to prosodic attributes, such as position of word in phrase (PinP), and position of syllable in word (PinW). Figure 3 displays the MDS plane of the KLD matrix for attribute PinP. It is observed that PinS elements are roughly grouped into 4 parts,

1. Head of phrase (PinS 3-1, 3-2).
2. Middle of phrase (PinS 1-1, 1-2, 2-1, 2-2).
3. Tail of phrase (PinS 1-3, 2-3).
4. PinS 3-3.

We intentionally separate PinS 3-3 from others, because it behaves in between head of phrase and tail of phrase, as is confirmed in Figure 4.

Further, by combining attributes PinS and PinW into a compound attribute, we could investigate the interaction between these two attributes. Figure 4 displays the MDS plane of the KLD matrix for attribute PinP*PinW. Symbols are expressed in the form of [PinS]:[PinW]. We observed that attribute PinW gains more priority in grouping elements than attribute PinP. Inside each PinW group, the group structure of PinS elements is generally maintained.

### 4.5. Evaluation for numeric contexts

In this section, we examined the characteristics of KLD with respect to numeric attributes, such as the number of phones in syllable (Phns), and the number of syllables in word (Syls). Frankly, if elements of a numeric attribute are of a limited set, there is nothing special in calculating KLD between these elements. What we emphasize here is that, though there exists an apparent metric for numeric attribute, KLD achieves a more reasonable one which agrees with their difference in acoustic characteristics. Here we projected elements on a one-dimensional space, Figure 5 for attribute Phns, and Figure 6 for attribute Syls. It shows that the elements keep the same order in the line as their values suggest, however they are not placed at as equal intervals. As values increase, their deltas,

on the whole, gradually decrease. This conforms to our knowledge on these attributes. As for attribute Phns, multiple-phone syllables typically sound different from monophone syllable, and the more phones in a syllable, the less increments of the effect they take in acoustic characteristics. The thing works the same for attribute Syls.

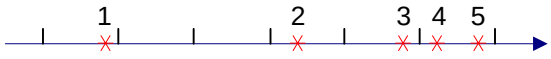


Figure 5. *MDS graph of the KLD matrix for attribute Phns.*

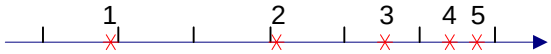


Figure 6. *MDS graph of the KLD matrix for attribute Syls.*

### 4.6. Subjective evaluation

In this section we evaluate the applicability of KLD as target cost in a task of speech synthesis. In our previous work on English TTS [15], the target cost consists of differences in phonetic and prosodic contexts, and the concatenation cost takes binary values: 0 when two segments to be concatenated are succeeding segments in the recorded speech, and 1 otherwise. Values in the cost function were perceptually tuned by language experts.

In the experiment, we substituted the manually-tuned target sub-costs with the KLD-based ones. Weights for each sub-cost were not studied in the paper and kept the same as original.

We did a preference test to compare the performance of KLD-based target cost with the original manually-tuned one. 30 sentences were synthesized as test stimuli based on a unit database of 5000 utterances. 8 subjects participated in the test and they were forced to choose one from each pair which sounds more natural.

The result for the preference test is given in Table 1. It shows that the synthetic speech obtained with the proposed KLD-based cost sounds slightly better than that with the manually-tuned costs.

Table 1: *Preference ratio for unit selection systems using KLD-based target cost and manually-tuned one.*

| | Manually-tuned | KLD-based |
|---|---|---|
| Pref. ratio | 45.3% | 54.7% |

## 5. CONCLUSION

In this paper, we employed KLD between context-dependent HMMs as the target cost in unit selection TTS systems. KLD between MSD-HMMs was presented. Also, we demonstrated that the KLD measure offers a statistically meaningful way to study the acoustic characteristics of speech-related attributes. With the help of MDS, we can visualize the underlining relations of elements from their KLD matrix. Perceptual experiments showed that the TTS system with the KLD-based target cost sounds slightly better than one with the manually-tuned.

Future works include examining the KLD measure on other speech-related attributes, such as part of speech. At this point, we lack the ability to optimize the weights for sub-costs. They may influence the voice quality more than sub-costs. We will investigate how to jointly optimize target sub-costs, weights of target cost, and even the concatenation cost.

## 6. References


[1] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", In *Proc. ICASSP 1996*, Atlanta, Georgia, 1996.

[2] A. Black and N. Campbell, "Optimising Selection of Units from Speech Databases for Concatenative Synthesis," in *Proc. Eurospeech 1995*, Madrid, Spain, 1995.

[3] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing Sub-cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis", in *Proc. ICASSP 2004*, Montreal, 2004.

[4] Y. Stylianou and A. K. Syrdal, "Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis", In *Proc. ICASSP 2001*, Salt Lake City, 2001.

[5] H. Peng, Y. Zhao and M. Chu, "Perpetually Optimizing the Cost Function for Unit Selection in a TTS System with One Single Run of MOS Evaluation," in *Proc. ICSLP 2002*, Denver, 2002.

[6] P. Taylor, "The Target Cost Formulation in Unit Selection Speech Synthesis", in *Proc. of Interspeech 2006*, Pittsburgh, 2006.

[7] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley Interscience, New York, NY, 1991.

[8] P. Liu, F. K. Soong and J.-L. Zhou, "Divergence-Based Similarity Measure for Spoken Document Retrieval", In *Proc. ICASSP 2007*, Hawaii, 2007.

[9] J. Goldberger, "An Efficient Image Similarity Measure Based on Approximations of KL-Divergence between Two Gaussian Mixtures", in *Proc. ICCV 2003*, Nice, France, 2003.

[10] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Multi-Space Probability Distribution HMM", IEICE Trans. Information and Systems, vol.E85-D, no.3, pp.455-464, Mar. 2002

[11] K. Tokuda, H. Zen, and A. Black, "An HMM-based Approach to Multilingual Speech Synthesis", in *Text to Speech Synthesis: New Paradigms and Advances*, pp. 135-153. Prentice Hall, 2004

[12] J. Odell, D. Ollason, P. Woodland, S. Young and J. Jansen, The HTK Book for HTK V3.0, Cambridge University Press, Cambridge, 2001.

[13] M. N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and hidden Markov Models", in *IEEE Signal Processing Letters*, Apr. 2003.

[14] F.W. Young, R.M. Hamer, Theory and Applications of Multidimensional Scaling, Eribaum Associates, Hillsdale, NJ, 1994.

[15] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, "Microsoft Mulan - a Bilingual TTS system," in *Proc. ICASSP 2003*, Hong Kong, 2003.

[16] J. Yang, Z. Zhao, Y. Jiang, G. Hu, and X. Wo, "Multi-Tier Non-Uniform Unit Selection for Corpus-Based Speech Synthesis", in *Proc. Blizzard Challenge 2006*, Pittsburgh, 2006.

[17] American English Phoneme Representation, in *Microsoft Speech SDK Version 5.1*, http://msdn.microsoft.com

[18] M. Beckman and J. Hirschberg. The ToBI Annotation Conventions, Ohio State University, Columbus, 1994