MEASURING TARGET COST IN UNIT SELECTION WITH KL-DIVERGENCE BETWEEN CONTEXT-DEPENDENT HMMS

Yong Zhao, Peng Liu, Yusheng Li, Yining Chen and Min Chu

Microsoft Research Asia, Beijing, China, 100080
{yzhao, pengliu, yushli, ynchen, minchu}@microsoft.com

ABSTRACT

This paper proposes a new approach for measuring the target cost in unit selection, where the difference between the target and candidate units is estimated by the Kullback-Leibler Divergence (KLD) between the context-dependent Hidden Markov Models (HMM). In order to model the left/right phonetic context, biphone models are generated by merging regular tri-phone HMMs sharing the same left/right phonetic context. To characterize prosodic contexts, various sets of prosody-sensitive monophone HMMs are trained. KLDs between these context models are calculated as the replacement cost between the contexts. Perceptual experiments show that the resulting synthesized speech sounds slightly better than those with the manually-tuned costs. An important advantage is that the proposed method can be conveniently applied to new corpora or languages without the need of collecting perceptual data.

1. INTRODUCTION

In unit-selection text-to-speech (TTS) systems, the task of unit selection is implemented by finding a sequence of database units that minimize the cost function. The cost function measures the distortion of the synthesized utterance, and is a summation of two sub-cost functions: a target cost, which describes the difference between the target segment and the candidate segment, and a concatenation cost, which reflects the smoothness of the concatenation between selected segments. Defining a cost function that can reflect concatenation unnaturalness just as a human listener might perceive it is significant for speech quality, and not a trivial task, since a number of factors are crucial for speech quality and their functionalities and interactions are not fully studied. Furthermore, it is difficult to evaluate the validity of a change in the cost calculation, since it normally improves the speech quality in some cases, but hurts the quality in other cases.

Various techniques have been proposed to optimize parameters in the cost function. CHATR system [1] used an automatic scheme for optimizing weights in the cost function by minimizing an objective measure between the reference sentence and the synthesized voice. A number of researchers have investigated the concatenation cost function by means of increasing the correlation between spectral distances and perceptual discontinuities [2][3].

In our previous work [4][5], the cost function was optimized by maximizing the correlation between the objective cost and the Mean Opinion Score (MOS). A set of synthesized utterances were scaled in MOS, and meanwhile the unit sequences composing these utterances were saved. By fixing the unit sequences, the objective cost can be recalculated with any changes in the cost function, and its correlation to the MOS can serve as an objective measure for examining the change validity. Finally, the optimized cost function has been significantly improved in its correlation to the MOS.

One of the shortcomings in the above researches is that the cost function must be optimized with respect to perceptual scores, which are difficult to collect. Thus, the investigated parameters are generally restricted within a limited range, such as joint smoothness on particular phone segments. Also, the optimization process is inconvenient to duplicate on new speech corpora, or languages.

In this paper, we apply KLD to estimate the target cost function, i.e. the mismatch cost between a target and candidate unit is calculated using the distances between the acoustic models that statistically represent these units. The proposed method has the advantage of applying to new corpora or languages without the need of collecting perceptual data.

This paper is organized as follows. Section 2 introduces the concept of KLD. The target cost used in our Mulan TTS system is summarized in Section 3. The method for modeling units in different contexts and calculating the target costs is explained in Section 4. The experiments and the perceptual evaluation results are presented in Section 5, and conclusions in Section 6.

2. KL-DIVERGENCE

The KLD [7][8] is a measure of the dissimilarity between two probabilistic models. If M and $\tilde{M}$ represent two

continuous distributions of feature x, the KLD between them is defined as:

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

For HMMs, the probability function is complex and there exists no simple closed form expression for calculating the KLD between HMM models. In [9], we proposed an effective algorithm for measuring KLD between two GMM-based HMMs.

The measure has been successfully applied to various model-based applications ranging from distortion measure to model clustering. Especially, in concatenative speech synthesis [2][3], KLD has been successfully used to calculate the spectral distance between two concatenated speech segments across the boundary, where spectral envelopes at the boundary are viewed as the probabilistic distributions in Eq. 1.

In this paper, KLD between context dependent HMMs is proposed as the measurement of target cost between a target and candidate unit, which are represented by their corresponding context dependent models, respectively.

3. THE TARGET COST IN MULAN TTS SYSTEM

In most concatenative TTS systems, all tokens of a unit are first clustered by their phonetic contexts, and then pruned by their distances from the core of the cluster or by their HMM scores. In these systems, prosodic features are used to select from several tokens within the same cluster. When the prosodic features of the selected unit do not match their predicted target, they will be scaled with signal processing methods. However, the synthesized speech often sounds unnatural. Typically, repetitious and monotonous prosody patterns are perceived, since natural variations in prosody of human speech are replaced with the most frequently used patterns.

In our previous work on the bilingual TTS system, Mulan [6], a prosody-constrained unit selection algorithm was adopted. The features used for searching the best token sequence in the unit selection phase incorporate two types: phonetic context features and prosodic context features. The phonetic context features are composed of Left Phonetic Context (LPhC) and Right Phonetic Context (RPhC), and the prosodic context features consist of Position in Phrase (PinP), Position in Word (PinW), Position in Syllable (PinS), Accent Level in Word (AinW) and Emphasis Level in Phrase (EinP). All of these features are discrete. Continuous prosodic features, like duration, power, and pitch, are excluded from the system, since it often produces monotonous prosodic patterns.

The cost function is defined as a weighted sum of the target cost and the concatenation cost. The target cost is the weight sum of distances in prosodic constraints and phonetic constraints; the concatenation cost takes binary values: 0 when the two segments to be concatenated are succeeding segments in the recorded speech and 1 otherwise.

Since each feature takes categorical values, which denote a context class, the distances between context classes of a candidate and target unit can be looked up in a distance table. The values in the table can be assigned by human experts. The classes that are far apart in perception should receive a high cost value, and vise versa.

Though the cost function used in [4][5] is proven to replicate, to a great extent, the perceptual behavior of human beings, it may not be optimal, even becoming worse for new speakers or new languages. Furthermore, in the current cost function, all items are assumed independent from each other without any deep study. In the next section, a method for automatically generating distance tables is introduced and the method can be used to generate phone dependent distance tables as well.

4. MEASURING TARGET COST WITH KLD BETWEEN CONTEXT DEPENDENT HMMS

The target cost takes into account the compatibility between the target and the candidate unit. Let $t = [t_1, t_2, \cdots, t_J]$ and $u = [u_1, u_2, \cdots, u_J]$ denote the target and candidate feature vectors, respectively. The original target cost is defined as:

$$C^t(t,u) = \sum_{j=1}^{J} w_j^t C_j^t(t_j, u_j) \quad (2)$$

where $C_j^t, j = 1,2,\cdots J$, is the sub-cost for the j^{th} element, and is weighted by w_j^t.

The sub-costs for categorical features can automatically be estimated by acoustically modeling the context classes of the feature. First, we build acoustic HMMs from the speech database to represent the context classes for each feature element; then the distances between the probabilities of these acoustic models are calculated as the mismatch between the corresponding context classes.

KL-Divergence is a sound measure of the dissimilarity between two probabilistic models, so it is intuitively adopted in our study. The target cost can be rewritten in terms of KL-Divergence as given in Eq. 3:

$$C^t(t,u) = \sum_{j=1}^{J} w_j^t D(T_j \| U_j) \quad (3)$$

where T_j and U_j denote the target and candidate models corresponding to unit feature t_j and u_j, respectively. A schematic diagram of measuring the target cost for unit t_i and u_j with KLD between HMMs is shown Figure 1.

The key problem in the KLD-based target cost estimation is how to build reliable context-feature HMMs,

which essentially characterize the addressed context classes, while removing the influence of other features.

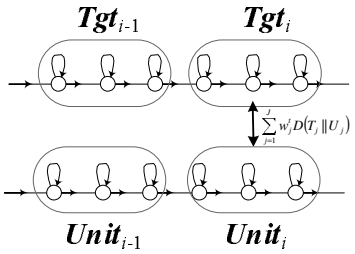

Figure 1: Schematic diagram of measuring the target cost with KLD between HMMs.

4.1. Phonetic target costs

The phonetic target cost consists of sub-costs for the Left Phone Context (LPhC) and the Right Phone Context (RPhC). For example, when selecting a unit /aw/ for the target phone sequence /m aw/, a unit /aw/ following a /m/ is desired; yet if only /aw/'s following other phones are available, we need a measurement that can help us to rank the similarity of LPhC of the available /aw/'s.

We propose a left biphone model to represent the LPhC. The models are estimated from the regular tri-phone HMMs [10], particularly accounting for the discrimination in the left phonetic context. Let *l-c+r* denote a triphone model, where *l*, *c*, and *r* are the left phone, center phone, and right phone, respectively. When our focus is on the LPhC of *c*, all triphone models with center phone *c*, the specified left phone *l* and whatever right phone are extracted and merged into a left biphone model *l-c* for *c*. As a result, the left biphone models are independent from their right context, i.e. the states on the right half of the model should have little discriminating information about the right phone context and the states on the left half of the model preserve the discrimination between left phones. The KLD between left biphone models *l-c* and *k-c* is calculated with the algorithm depicted in [9]. To express the mismatch between LPhC *l* and *k* for unit *c*, the KLD is further normalized with a linear scaling function into a fixed range, like from 0 to 1, so that the weight of the sub-cost for this feature can remain somewhat stable.

In such a way, the distances for any left or right phone mismatch can be represented by the normalized KLD of left/right biphone models of a unit. It should be pointed out that the distance tables obtained with the proposed method are unit dependent, i.e. if we define 40 phones for English, we will obtain 40 distance tables.

4.2. Prosodic target costs

The prosodic target costs consist of the sub-costs for Position in Phrase (PinP), Position in Word (PinW), Position in Syllable (PinS), Accent Level in Word (AinW) and Emphasis Level in Phrase (EinP), etc.

We present a scheme to build separate sets of prosody-sensitive monophone HMMs to represent different prosodic categories. Here, we use PinW as an example. It has 4 categorical values: at Head, Middle, Tail of a word, or a Mono-syllable word. To model PinW context, the base monophone HMMs are first trained, then each base phone is expanded into 4 PinW sensitive HMMs, i.e. the expanded phone set takes into account PinW with the form of *c:x*, where *x* is phone *c*'s PinW label. For example, the word 'robot' with pronunciation /r ow – b ax t/ is composed of two syllables, where the first syllable is with PinW Head, and the second with Tail, thus the phones are expanded with the form as /r:h ow:h – b:t ax:t t:t/, where *h* stands for Head, *t* for Tail. The normalized KLD between models $c:x_1$ and $c:x_2$ represents the distance between PinW x_1 and x_2 for unit *c*.

All the prosodic target costs can be calculated by creating models of mono-phones extended with prosodic labels, and comparing the models with KLD. Likewise, the distance tables are unit dependent. Table 1 shows an example for the model expressions of a unit sequence.

Table1: *An example for the model expressions on various contexts. (The input sentence is 'All together'; In Pronunciation layer, 1 means stressed syllable and '-' means syllable boundary; In Position in Word layer, h, i, t and m denote Head, Middle, Tail and Mono; In Position in Syllable layer, o, n, c and m, stand for Onset, Nucleus, Coda, and Mono, respectively; In Accent in Word layer, 1 denotes stressed, 0 unstressed.)*

Text:	All		together.						
Pronunciation:	/ao 1 l/		/t ax - g eh 1 - dh ax r/						
Target unit string:	ao	l	t	ax	g	eh	dh	ax	r
Left Phone Context:	sil-ao	ao-l	l-t	t-ax	ax-g	g-eh	eh-dh	dh-ax	ax-r
Right Phone Context:	ao+l	l+t	t+ax	ax+g	g+eh	eh+dh	dh+ax	ax+r	r+sil
Position in Word:	ao:m	l:m	t:h	ax:h	g:i	eh:i	dh:t	ax:t	r:t
Position in Syllable:	ao:o	l:c	t:o	ax:c	g:o	eh:c	dh:o	ax:n	r:c
Accent in Word:	ao:1	l:1	t:0	ax:0	g:1	eh:1	dh:0	ax:0	r:0
.....									

5. EXPERIMENTS AND RESULTS

5.1. Experimental setup

The Microsoft Mulan English TTS speech corpus is used to evaluate the performance of the proposed method. The

corpus contains about 6000 sentences, which were recorded by a professional speaker and annotated with symbolic prosodic labels, like break level, stress, emphasis, etc.

Cost tables for each feature component are calculated separately. Context-dependent HMMs are trained for each context separately. As for the phonetic target cost, 3-state HMMs are trained for all triphones, where 39-dimensional MFCC features are extracted and 8 Gaussian Mixtures with diagonal covariance matrices are computed for each state. Then the left/right biphone HMMs are generated by merging triphone models sharing the same phonetic contexts. Similarly, prosody-sensitive monophone HMMs are trained for calculating the prosodic target costs. KLDs between these models are computed as the replacement cost between the target and candidate units.

The weights for each component are not studied in this paper. At this point, these weights are maintained as the original manually-tuned weights.

5.2. Subjective evaluation

700 sentences are synthesized with both the KLD-based cost tables and the manually-tuned tables. They are generated in two conditions: using only 1500 sentences from the speech database (denoted as DB1) and using the full speech database (denoted as DB2). Among them, 60 pairs of sentences with the maximal difference ratio in their selected unit sequence are used for the listening test. 8 subjects participate in the test and they are forced to choose one from each pair which sounds more natural.

The result for the preference test is given in Table 2. It shows that the synthetic speech obtained with the proposed KLD-based cost sounds slightly better than that with the manually-tuned costs. It is also observed that the preference on DB2 with the KLD-based cost performs a little better than that on DB1 (also, the difference between the two sets on DB2 is significant, $P<0.00001$). One possible reason is that DB2 contains many more candidate units, which enable the subtle difference between competing context features to be exposed.

Table 2: *Preference test results. (A: manually-tuned; B: proposed method)*

	Prefer A	Prefer B
DB1	47.2%	52.8%
DB2	45.3%	54.7%

6. CONCLUSION

In the paper, we proposed a novel method for measuring the target cost in unit selection with KLD between context-dependent Hidden Markov Models (HMM). Various context-dependent HMM models are trained to represent units in different phonetic and prosodic contexts. KLD between pairs of these models are used as the approximation of the replacement cost between the two contexts. Perceptual experiments show that the resulted synthesized speech sounds slightly better than those synthesized with the manually-tuned costs. More importantly, the proposed method is easy to apply to any new corpora or languages, since the cost weights are trained without involving perceptual scores.

Future work will be directed towards integrating the concatenation cost function and the component weights into the proposed scheme. Furthermore, we will investigate the capability of the KLD-based cost tables for other languages.

7. REFERENCES

[1] A. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proc. of Eurospeech1995*, Madrid, Spain, 1995.

[2] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. of ICASSP2001*, Salt Lake City, 2001.

[3] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," in *Proc. of ICSLP2002*, Denver, 2002.

[4] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech", in *Proc. of Eurospeech2001*, Aalborg, 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. of ICSLP2002*, Denver, 2002.

[6] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, "Microsoft Mulan - a bilingual TTS system," in *Proc. of ICASSP2003*, Hong Kong, 2003.

[7] S. Kullback and R. A. Leibler, "On Information and Sufficiency", in Ann. Math. Stat., 22: 79-86, 1951.

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

[9] Peng Liu, Frank K. Soong, Jian-Lai Zhou, "Effective Estimation of Kullback-Leibler Divergence between Speech Models", Microsoft Research Asia, Technical Report, 2005.

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