Automatic Error Detection for Unit Selection Speech Synthesis Using Log Likelihood Ratio based SVM Classifier

Heng Lu, Zhen-Hua Ling, Si Wei, Li-Rong Dai, Ren-Hua Wang

iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China

luhenglh@mail.ustc.edu.cn

Abstract

This paper proposes a method to detect the errors in synthetic speech of a unit selection speech synthesis system automatically using log likelihood ratio and support vector machine (SVM). For SVM training, a set of synthetic speech are firstly generated by a given speech synthesis system and their synthetic errors are labeled by manually annotating the segments that sound unnatural. Then, two context-dependent acoustic models are trained using the natural and unnatural segments of labeled synthetic speech respectively. The log likelihood ratio of acoustic features between these two models is adopted to train the SVM classifier for error detection. Experimental results show the proposed method is effective in detecting the errors of pitch contour within a word for a Mandarin speech synthesis system. The proposed SVM method using log likelihood ratio between context-dependent acoustic models outperforms the SVM classifier trained on acoustic features directly.

Index Terms: Speech synthesis, error detection, log likelihood ratio, support vector machine

1. Introduction

Speech synthesis has been under development for many years. Nowadays, there are mainly two kinds of speech synthesis methods. One is the HMM-based parametric speech synthesis [1] and the other is the unit-selection and waveform concatenation [2, 3]. Both the two methods have their advantages and disadvantages. HMM-based parametric speech synthesis can generate synthesis speech using a relatively small database, and it performs robustly and rarely make serious errors. However, the average effect of statistical modeling makes its synthetic speech sound muffled. Unit selection and waveform concatenation method can generate synthetic speech with high naturalness if the speech corpus is large enough. However, when the speech corpus can not provide satisfactory coverage for contextual factors there often occur some serious errors in the synthetic speech which degrades the robustness of synthesis system significantly.

Besides, the aim of a speech synthesis system is to synthesize speech as natural as the ones spoken in everyday life. And in order to evaluate the naturalness of synthetic speech, the Mean Opinion Score(MOS) which is a subjective perception score given by listeners is utilized. Unfortunately, none of the two main-stream speech synthesis methods involves any subjective evaluation on synthetic speech into the procedure of system construction. HMM-based parametric speech synthesis aims to generate speech parameters which lead to the maximum likelihood given the trained HMM models [4]. And in unit selection speech synthesis systems, cost functions composed of the distances on acoustic parameters or contextual descriptions between the candidate and target units are minimized to select the best unit sequence for waveform concatenation. Neither the likelihood nor the cost functions can present the subjective human perception of synthetic speech explicitly.

Therefore, this paper studies how to model the human perception on synthetic speech and focuses on detecting the synthetic errors for a unit selection synthesis system. Once the errors in the synthetic speech can be detected automatically, it can be used as a criterion to guide the selection of optimal unit sequence to unify the objective unit selection criterion and the subjective performance measurement. Similar error detection problem has been studied for many years in computer assisted language learning (CALL) systems [5–7], which aim to detect the pronunciation errors of a speaker when learning languages. In [8], a pronunciation space model (PSM) based SVM method was introduced to detect the pronounced errors. In this method, the distribution of acoustic features was divided into several spaces according to the correctness of pronunciation and an acoustic model was trained for each space. Posterior probabilities of the pronunciation spaces were used as input features to train an SVM for pronunciation error detection. This PSM-SVM method outperformed the traditional SVM method where acoustic feature were directly utilized to train the classifier. In [9], a performance evaluation method for synthetic speech was proposed, where qualitative and quantitative features were extracted to train a SVM directly to evaluate the correctness of pitch accent in a Japanese unit selection speech synthesis system.

In this paper, a log likelihood ratio based SVM method is proposed to detect the synthetic errors of a unit selection synthesis speech system. Similar to PSM, we divide the acoustic features of synthetic speech into two spaces, which represent *natural* and *unnatural* pronunciation given by a speech synthesis system respectively. Context-dependent acoustic models are trained for these two spaces separately. For each segment of synthetic speech in the training set, log likelihood ratios of acoustic features between these two models are used as the input vector and its naturalness given by subjective evaluation is used as the output to train an SVM classifier. Experimental result reveals the proposed method outperforms the SVM method using acoustic features directly.

The rest of this papers is organized as follows. Section 2 describes the proposed method in detail. Section 3 introduces our implementation using a Mandarin speech synthesis system and the experimental results. Conclusions and future work are presented in Section 4.

2. Method

Our proposed method uses a log likelihood ratio based SVM classifier to detect the errors, i.e., unnatural segments, in the

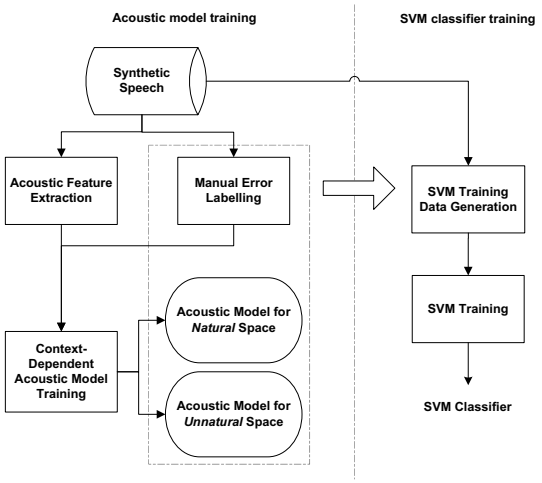

Figure 1: Flowchart of SVM training for synthetic error detection.

synthetic speech. The flowchart of SVM training procedure is shown in Fig. 1, which consists of three main steps: 1) training set construction by manual error labelling on synthetic speech; 2) acoustic feature extraction and context-dependent acoustic model training; and 3) SVM classifier training. These three steps will be introduced in turn next.

2.1. Synthetic error definition and annotation

We define "synthetic error" as the segment in synthetic speech that is perceived to be unnatural by listener. Compared with the errors people make during pronunciation, such as mispronouncing one phoneme with another, the errors in synthetic speech may be a drastic pitch rising or a click sound at the concatenation point, which are more complex and difficult for detection. There are many factors that can cause the synthetic errors, including articulation, pitch accent, duration and power, as mentioned in [9]. Our preliminary subjective evaluating results show that inappropriate pitch contour is the most significant one among these factors for a Mandarin unit-selection speech synthesis system, which causes more than 70% of all synthetic errors.

In order to train the SVM error detector, a training set of synthetic speech with manual labelling results of synthetic errors is a precondition. For a given speech synthesis system, a set of phonetically balanced texts are firstly synthesized. These synthetic sentences are presented to the labelers for error annotation. In our work on Mandarin speech synthesis, prosodic word is selected as the basic segment in labelling because the annotation using smaller unit, such as syllable or phone, is proved to be difficult and inconsistent among labelers in our experiments. One word is annotated to be either *natural* or *unnatural* according to one factor of human perception, such as articulation, pitch, or duration. Here, the synthetic errors caused by different factors are labelled independently because we can extract proper features in the following acoustic model training to detect the corresponding errors.

2.2. Acoustic model training

As discussed above, the extraction of acoustic features depends on the type of synthetic errors we want to detect. For example, the acoustic features representing spectral characters, such as mel-frequency cepstral coefficients (MFCC) or line spectral pairs (LSP) [10], can be used to detect the improper articulation in synthetic speech. Besides, F0 is an important feature to detect the synthetic error of pitch accent. For Mandarin which is a tonal language, many pitch errors are caused by the mismatch of pitch contour between adjacent syllables. Therefore, the long-term F0 features which describe the relationship of F0 trajectory among nearby syllables should be included in the feature vectors for acoustic model training.

For acoustic model training, the acoustic features of synthetic speech in the training set are divided into a *natural* space and an *unnatural* space according to the manual labelling results. Context-dependent acoustic models are trained for these two spaces independently. The basic unit of acoustic model, such as phone or syllable, and the model structure, such as Gaussian mixture model (GMM) or hidden Markov model (HMM), can be set according the characteristics of extracted acoustic features. For each space, after training the fully context-dependent acoustic model, a decision tree based model clustering is conducted using a well designed question set on linguistical contexts to avoid the data sparsity problem and predict model parameters for the contexts out of training set [11].

2.3. SVM training

In accordance with the manual error labelling introduced in Section 2.1, prosodic word is taken as the basic segment in the SVM classifier training process. Assuming a prosodic word consists of N syllables[1] and the extracted acoustic feature has M dimensions, an $N * M$ dimensional vector $\mathbf{\Omega}$ is composed as

$$\mathbf{\Omega} = \left[\omega_1^1, ..., \omega_1^M, \omega_2^1, ..., \omega_2^M, ..., \omega_n^m, ..., \omega_N^1, ..., \omega_N^M\right]^\top, \quad (1)$$

and

$$\omega_n^m = \log P_m(\boldsymbol{o}_n|\lambda_N, c_n) - \log P_m(\boldsymbol{o}_n|\lambda_U, c_n), \quad (2)$$

where $n = 1, ..., N; m = 1, ..., M$; $\boldsymbol{o}_n$ stands for the acoustic features belong to the n-th syllable in the word; λ_N and λ_U denote the trained acoustic models for *natural* and *unnatural* space respectively; c_n means the context information of syllable n; $P_m(\boldsymbol{o}|\lambda, c)$ indicates the component of likelihood calculated for the m-th dimension of feature $\boldsymbol{o}$ given model λ and context c.

In this way, for each prosodic word in the training set with manual error labelling result s, an instance of $(\mathbf{\Omega}, s)$ can be generated for the SVM training, where $\mathbf{\Omega}$ is the input vector and s is output of SVM classifier. Since a prosodic word may contain different number of syllables, this lead to a variable vector size of $\mathbf{\Omega}$. Thus,we construct a group of SVM detectors for the prosodic words with different number of syllables.

2.4. Synthetic error detection using SVM

After the SVM is trained, it can detect the errors within any input synthetic speech automatically. For each prosodic word in the synthetic speech, acoustic features are firstly extracted and an input vector is then calculated as (1) according to the context information of current word. Finally, this vector is sent into the trained SVM classifier to judge whether this word contains synthetic errors.

[1] Syllable is assumed to be the basic unit of trained acoustic models here to facilitate the introduction.

3. Implementation

3.1. The speech synthesis system for error detection

A speech synthesis system provided by iFLYTEK company for GPS navigation application was used in our experiments for synthetic error detection. The language of the system was Mandarin and its task was to synthesize the name of places for GPS navigation. This system was constructed following the HMM-based unit selection and waveform concatenation approach [12] and using a speech database of 6,000 short sentences broadcasted by a female speaker.

The construction of this HMM-based unit-selection speech synthesis system consists of two main parts, the HMM model training part [13] and the speech synthesis part. In the HMM model training part, acoustic parameters are extracted from the speech database, including spectral and prosody features. Then the spectral features are modeled by a continuous probability HMM and the F0 features are modeled by a multi-space probability HMM (MSD-HMM) [14]. As there are enormous combinations of context features, minimum description length (MDL) [11] based HMM model clustering is conducted to avoid data sparse problem and to predict models for the text to be synthesized. Phone duration model is also trained to model the duration of phone. Apart from the acoustic models mentioned above, spectral and F0 concatenating models are introduced to measure the smoothness at the concatenated phone boundaries in the synthesized speech. In the synthesis part, five statistical models, including the spectrum model, F0 model, phone duration model, concatenating spectrum model and concatenating F0 model are employed to guide the selection of the most suitable unit sequence from the database under maximum likelihood criterion. Then selected phone units are concatenated to generate the speech waves. In the unit selection process, Kullback-Leibler divergence (KLD) [15] between the model of the candidate unit and the target model is used to conduct the unit pre-selection in order to reduce computational cost in the Viterbi unit selection process.

3.2. Pitch error annotation

We focus on detecting the synthetic errors caused by unnatural pitch patterns in our experiments. As we know, Mandarin is a tonal language and syllables with different tone have completely different meanings. There are four basic tones (high level, high rising, low falling-rising, and high falling) and one neutral tone in Mandarin. Pitch errors in the synthetic speech of Mandarin can be one tone misperceived as another, or the unnatural combination of pitch contours for adjacent syllables.

A text set consisting of 169,492 names of places was used in our experiments. Theses texts were synthesized using the speech synthesis system introduced in Section 3.1 and the synthetic results were manually labelled as discussed in Section 2.1. Here, each sentence consisted of only one prosodic word for the text of place names. 54,355 prosodic words were labelled with synthetic errors on pitch perception, while the other 11,7397 were labelled as natural ones. And 90% of these synthetic sentences were randomly selected to construct a training set, while the remaining ones were used as a test set.

3.3. Acoustic model training

For each syllable in the synthetic speech of training set, a F0 trajectory was extracted and equally split into 5 parts. Let $p_{n,i}$ denote the average log F0 of the i-th part of the n-th syllable in a prosodic word; $\bar{p}_s$, $\bar{p}_w$ and $\bar{p}_p$ denote the average log F0 of current syllable, word and phrase respectively. Then, an acoustic feature vector of 13 dimensions was defined for syllable n as shown in Table 1. Unlike the frame-based F0 features extracted for HMM training in the speech synthesis system construction, some F0 features that can present the long-term characters of pitch contour, such as the difference between the average log F0 of adjacent syllables, were adopted here for the error detection system training.

Dimension	Acoustic feature definition
$1 \sim 5$	$p_{n,i}, i = \{1,2,3,4,5\}$
6	$p_{n,1} - p_{n-1,5}$
7	$p_{n,5} - p_{n+1,1}$
8	$p_{n,1} - \bar{p}_w$
9	$p_{n,5} - \bar{p}_w$
10	$\bar{p}_s - \bar{p}_w$
11	$p_{n,1} - \bar{p}_p$
12	$p_{n,5} - \bar{p}_p$
13	$\bar{p}_s - \bar{p}_p$

Table 1: Definition of acoustic feature vector for syllable n in a prosodic word. $p_{n,i}$ denote the average log F0 of the i-th part of the n-th syllable; $\bar{p}_s$, $\bar{p}_w$ and $\bar{p}_p$ denote the average log F0 of current syllable, word and phrase respectively.

Dataset	# of natural words	# of errors
Training Set	31114	15770
Test Set	3458	1753

Table 2: Number of 3-syllable word in the training set and the test set.

Using the acoustic feature vectors extracted for all syllables in the training set, two context-dependent acoustic models were trained to present the *natural* and *unnatural* space independently according to the manual annotation given in Section 3.2. Here, a single Gaussian distribution was adopted as the model structure and its covariance matrix was set to be diagonal.

3.4. SVM classifier training

A group of SVM classifiers were trained for the prosodic words with different number of syllables following the method described in Section 2.3. We take the SVM for 3-syllable word as an example in the following introduction. The numbers of 3-syllable word in the training set and test set are shown in Table 2. The N and M in Section 2.3 were set to 3 and 13 respectively.

Besides our proposed method, an SVM trained using acoustic feature vectors directly was also built as a comparison in evaluation. Instead of using log likelihood ratio as (2), the 13 dimensional F0 features in Table 1 were simply combined to produce the input vector $\mathbf{\Omega}$ for SVM training as

$$\mathbf{\Omega} = [o_{1,1}, ..., o_{1,M}, o_{2,1}, ..., o_{2,M}, ..., o_{N,1}, ..., o_{N,M}]^\top , \quad (3)$$

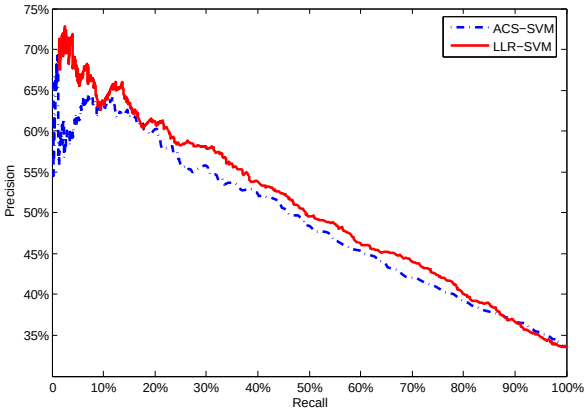

Figure 2: Precision-recall curve of SVM-based synthetic error detector on the test set for 3-syllable word. *ACS-SVM* denotes the SVM trained on acoustic features directly and *LLR-SVM* denotes our proposed method using log likelihood ratio.

where $o_{n,m}$ denotes the F0 feature of syllable n and dimension m; $N = 3$ and $M = 13$. The LibSVM toolkit [16] was employed to train the SVM classifiers and the radial basis function (RBF) kernel was used for both SVM systems.

3.5. Experimental results

By modifying the thresholds for the output of SVM classification probability, precision-recall curves for these two SVM systems were calculated and are shown in Fig. 2. Considering the number of natural and unnatural words in the test set as listed in Table 2, random detection would lead to a precision of about 33.6%. Fig. 2 shows that the performance of our proposed method is much better than the random detection. We can also see than *LLR-SVM* outperforms *ACS-SVM* because the context information was taken into account in the training of acoustic models and the calculation of log likelihood ratio. For example, a pattern of pitch contours may sound natural for a prosodic word with a given combination of tones or at a particular position in the sentence, whereas it may be unnatural when context information changes. Different from the *ACS-SVM* system, our proposed method can model such influence of contexts on the subjective perception.

4. Conclusions

This paper has presented a log likelihood ratio and SVM classifier based method to detect the synthetic errors for a unit selection speech synthesis system. The idea of pronunciation space model (PSM) is introduced into the training of SVM: the input feature vector is computed as the log likelihood ratio of acoustic features between the context-dependent statistical models presenting *natural* space and *unnatural* space; the results of manual error labelling on synthetic speech are used as the output for SVM training. Experimental results have shown that the proposed method can detect the pitch errors for a Mandarin speech synthesis system effectively and it outperforms the SVM classifier directly trained on acoustic features without using context information. Our future work includes improving the naturalness of a speech synthesis system by using the results of our proposed automatic error detection method as a feedback to refine the unit selection algorithm, and evaluating synthetic speech automatically using a quantitative score correlated with the MOS given by listeners.

5. Acknowledgements

This work was partially supported by the China Postdoctoral Science Foundation and the National Nature Science Foundation of China (Grant No. 60905010). The authors thank the research division of iFLYTEK company for providing the speech synthesis system used in our experiments and the manual error labelling results on the synthetic speech for system training.

6. References

[1] A. W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis”, in Proc.ICASSP, vol. 4, 2007, pp. 1229-1232.

[2] N. Iwahashi, N. Kaiki, and Y. Sagisaka, “Speech segment selection for concatenative synthesis based on spectral distortion minimization”, IEICE Trans. Fundamentals, vol.E76-A, no.11, pp.1942-1948, Nov. 1993.

[3] A.J. Hunt and A.W. Black, “Unit selection in a concatenative speech synthesis system using a large speech databas”, in Proc.ICASSP, Atlanta, USA, May 1996, pp.373-376.

[4] K. Tokuda, T. Kobayashi and S. Imai, “Speech parameter generation from HMM using dynamic features”, in Proc.ICASSP, pp.660C663, 1995.

[5] C. Cucchiarini, F. Wet, H. Strik, and L. Boves, “Assessment of Dutch pronunciation by means of automatic speech recognition technology”, in Proc.ICSLP, pp.1739-1742.

[6] H. Franco, L. Neumeyer, Y. Kim, O. Ronen,and H. Bratt, “Automatic detection of phone-level mispronunciation for language learning”, in Proc.of European Conference on Speech Communication and Technology, pp. 851-854.

[7] H. Franco, L. Neumeyer, V. Digalakis, O. Ronen, “Combination of machine scores for automatic grading of pronunciation quality”, in Speech Communication, pp.121-130.

[8] S. Wei and G-P. Hu, Y. Hu and R-H. Wang, “A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models”, in Speech Communication, vol 51, no 10, Oct. 2009, pp. 896-905

[9] A. Yoshida, H. Mizuno, and K. Mano, “Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis ”, in Proc.ICASSP, pp.4617-4620, 2008.

[10] I.V. McLoughlin, “Review: Line spectral pairs”, in Signal Processing, vol 88, no 3, March . 2008, pp. 448-467

[12] Z.-H. Ling, and R.-H. Wang, “HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion”, in Proc. ICASSP, 2007, pp. 1245-1248.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis ”, in Eurospeech, 1999, pp. 2347-2350

[14] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Markov models based on multi-space probability distribution for pitch pattern modeling ”, in ICASSP, 1999, pp. 229-232.

[11] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition ”, in J. Acoust. Soc. Japan(E), vol. 21, no. 2, pp. 79-86, 2000.

[15] S. Kullback and R. A. Leibler, “On information and sufficiency, ”, in Ann. Math. Stat., vol. 22, pp. 79-86, 1951.

[16] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/