TEMPLATE CONSTRAINED POSTERIOR FOR VERIFYING PHONE TRANSCRIPTIONS

Lijuan Wang[1], Tao Hu[2], and Frank Soong[1]

[1]Microsoft Research Asia, Beijing, China
[2]Information Security Engineering College, Shanghai Jiao Tong University, Shanghai, China
[1]{lijuanw, frankkps}@microsoft.com;[2]tao@sjtu.edu.cn

ABSTRACT

A new statistical confidence measure, Template Constrained Posterior (TCP), is proposed for verifying phone transcriptions of speech databases. Different from generalized posterior probability (GPP), TCP is computed by considering string hypotheses that bear a focused unit, e.g., phone with partially matched left and right contexts. Parameters used for TCP include context window length, partial matching ratio, KLD threshold for selecting confusable phones, and verification threshold. They are determined by minimizing verification errors in a development set. Evaluated on a test set which contains 52.1% sentence errors and 0.62% phone errors, TCP achieves 92% and 88% error hit rate in rejected sentences, when the corresponding acceptance ratios are set at 90% and 80%, respectively.

Index Terms—template constrained posterior, TCP, confidence measure

1. INTRODUCTION

Human-computer voice interaction via text-to-speech and speech recognition has been an intensive subject of research for many years. One significant issue in this field is that nearly all work must rely upon a well-annotated speech database. For example, text-to-speech synthesis relies upon the accuracy of annotated phonetic labels and corresponding contexts for selecting good acoustic units from a pre-recorded database. However, such a database must be thoroughly examined before it may be relied upon, in order to catch reading or pronunciation errors, transcription errors, incomplete pronunciation lists, and similar issues. Because of the importance and wide application of this issue, automated detection of error is highly desirable. Confidence is a useful measure for verifying speech transcription by assessing the reliability of a focused unit, such as a word, syllable, or phone.

A number of approaches for measuring confidence of speech transcriptions have been investigated [1-6]. They can be roughly classified into three major categories: 1) Feature based approaches that attempt to assess confidence based on selected features, such as word duration, part of speech, or word graph density, using trained classifiers; 2) Explicit model based approaches that use a candidate class model with competing models, and a likelihood ratio test; 3) Posterior probability approaches that attempt to estimate the posterior probability of a recognized entity, given all acoustic observations.

In this study we propose a new confidence measure, Template Constrained Posterior (TCP), for verifying transcription errors. The In the phone transcription verification, a *template* is constructed to compute phone level TCP, which considers not only the focused phone, but also the partially matched contexts before and after the focused phone. The template can be flexibly tailored to provide various granularities, from a finely defined context to loosely defined contexts. It effectively limits the hypothesis set used in calculating the posterior probability for a selected focus unit. Tests on an English TTS corpus show that TCP can effectively detect the erroneous sentences with 1~2 subtle phone errors, with 92% error hit rate when 10% sentences are rejected.

The rest of the paper is organized as follows. Section 2 introduces generalized posterior probability. Section 3 presents the new Template Constrained Posterior (TCP). Section 4 shows how to verify phone transcription using TCP and gives the experimental results. Section 5 draws the conclusions.

2. GENERALIZED POSTERIOR PROBABILITY

Generalized posterior probability (GPP) [1,2] is often used in speech transcription analysis as a confidence measure for verifying hypothesized entities at phone, syllable, or word levels. For a selected focus unit, e.g., a word, the acoustic probability and the linguistic probability of that word are compared against the total set of possible hypotheses to generate a ratio. The higher the calculated GPP, the more probable that the focus unit was correctly transcribed. Eq. 1, below, defines this relationship.

$$p(w \mid x_1^T) = \frac{\sum_{h \in H} p(h)}{\sum_{h \in R} p(h)}, \quad H \subset R \tag{1}$$

Let R represent the search space, which includes all possible string hypotheses for a given sequence of acoustic observations x_1^T. In practice, the search space R is usually reduced to a pruned space, for example a word graph. H, a subset of R, contains all string hypotheses that include/cover the focused word “w” by a given time range between starting and ending points. The posterior probability of “w” can be obtained by Equation 1, i.e., the sum of the probabilities of string hypotheses in H divided by the sum of probabilities of string hypotheses in R. Therefore, finding the right hypothesis subset H of R is a critical step in computing the posterior probability P(w| x_1^T) for verification. Eq. 2, below, provides an example equation for calculating generalized word posterior probability [2].

$$p([w;s,t] \mid x_1^T) = \sum_{\substack{N,[w;s,t]_1^N \\ \exists n,\ 1 \le n \le N \\ w = w_n \\ [s,t] \cap [s_n,t_n] \ne \phi}} \frac{\prod_{n=1}^{N} p^{\alpha}\left(x_{s_n}^{t_n} \mid w_n\right) \cdot p^{\beta}\left(w_n \mid w_1^N\right)}{p\left(x_1^T\right)} \tag{2}$$

GPP has several shortcomings. Firstly, as the graph, or search space, becomes richer, the probability of a competing unit increases. Secondly, selection of the correct hypothesis subset H is dependent upon the provided time frame; if the start or end time is inaccurate, e.g., due to deletion, substitution, or addition error in the neighborhood, the probability of the focused unit within the given time frame can be substantially altered.

3. TEMPLATE CONSTRAINED POSTERIOR

Through the use of templates, this study seeks to avoid the shortcomings of GPP. Use of templates allows a "sifting" of hypotheses; only those hypotheses which match both the focus unit and specified contexts are included in the search space, which leads to higher calculated probability results for the focus unit, and greater confidence. Moreover, since the templates are flexibly constructed, TCP can either be reduced to the traditional GPP, which considers only the focus unit, or be built upon a template of complex topology, where specific context for the focus word is defined.

It should be emphasized that for discussion convenience, this paper uses phone as the focus unit. However, the proposed method is also applicable to unit at other levels, such as syllable, word, phrase, sentence, etc.

3.1. Template and its variation

We denote a Template by a triple [$\mathcal{T}$; *R*; *s*, *t*]. Template $\mathcal{T}$ is a pattern composed of hypothesized units and metacharacters that can support regular expression syntax; R stands for the partial match ratio and ranges between 0 and 100%. This means the relevant path needn't 100% match the template. Partial match of R% against the template is acceptable. [s, t] defines the time frame constraint on the template.

As shown in Figure 1, Basic template $\mathcal{T}_1$ depicts the simplest type of template, ABCDE, where C is the focus unit, and AB and DE are the left and right context, respectively. Template $\mathcal{T}_2$, A*CDE, includes a wild-card, *, to indicate that the template does not care what appears in that particular position: A*CDE matches AACDE, AFCDE, or ACDE. Template $\mathcal{T}_3$, ABCφE, includes a blank, φ, to indicate a null in this position. Template $\mathcal{T}_4$, ABC?E, includes a question mark, ?, to indicate that the word which appears in this position has not been identified yet.

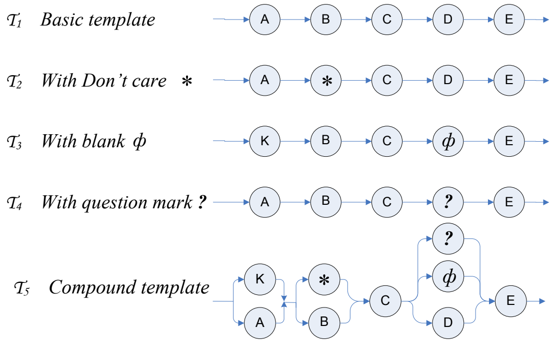
$\mathcal{T}_1$ Basic template
A
B
C
D
E
$\mathcal{T}_2$ With Don't care *
A
*
C
D
E
$\mathcal{T}_3$ With blank φ
K
B
C
φ
E
$\mathcal{T}_4$ With question mark ?
A
B
C
?
E
$\mathcal{T}_5$ Compound template
K
A
*
B
C
?
φ
D
E

Figure 1: *Illustration of templates.*

These basic templates can be combined to construct a compound template, such as template $\mathcal{T}_5$ depicted in Figure 1. With reference to compound template $\mathcal{T}_5$, a matching string hypothesis may include either A or K in the 1st position, include B or any element at the 2nd position, includes C at the center position, and so on. Depending upon the specified minimal matching constraint and whether some or all of these elements can be partially matched, the search space generated from compound template $\mathcal{T}_5$ may be substantially larger than that generated from a basic template.

3.2. Exemplary template for a focused phone

Eq. 3 gives a template example for a focused phone, which can be visualized in Figure 2, where p_k is the focused phone, $p_{k-L}\cdots p_k\cdots p_{k+L}$ is the phone string covering the 2*L* context phones before and after p_k. $\tilde{p}_i$ represents the confusable phone of p_i (k-L≤i≤k+L). With the help of regular expression, $\left[p_i\tilde{p}_i\right]$ matches either p_i or any confusable phones $\tilde{p}_i$. R is the partial match ratio among the *2L* context phones. [s, t] defines the time frame constraint of the template, i.e., *s* is the start time of p_{k-L} and *t* is the end time of p_{k+L}. The correct hypotheses set H for [T ; *R; s,t*], as defined in Eq. 1, is obtained by finding every string hypothesis that contains a subpath that R% partially matches the template and also overlaps the specified time interval [s, t].

$$\mathcal{T} = \left[p_{k-L}\tilde{p}_{k-L}\right]\cdots\left[p_{k-1}\tilde{p}_{k-1}\right]p_k\left[p_{k+1}\tilde{p}_{k+1}\right]\cdots\left[p_{k+L}\tilde{p}_{k+L}\right] \quad (3)$$

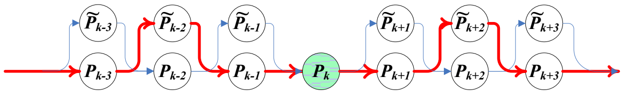

Figure 2: *Illustration of template for the focused phone* p_k.

In this study, the confusability between two phones is assessed by the Kullback-Leibler Divergence (KLD), which is a measure of the dissimilarity between two probabilistic models [8]. Given a threshold T_{KLD}, for two different phones, p_i and p_j, if $KLD(\lambda(p_i)\,\|\,\lambda(p_j)) \le T_{KLD}$, p_i is one of the confusable phones of p_j, and vice versa. For example, when T_{KLD} is set as 100, no confusable phone pair exists. When set as 300, "ih" has 3 confusable phones: "eh", "uh", and "ax". When T_{KLD} is set further larger, more confusable phones will appear.

3.3. Method of calculating TCP

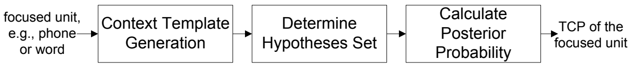
focused unit, e.g., phone or word
Context Template Generation
Determine Hypotheses Set
Calculate Posterior Probability
TCP of the focused unit

Figure 3: *A flowchart of calculating TCP.*.

Figure 3 depicts the flowchart of calculating template constrained posterior (TCP). Firstly, a focus unit is selected, which may be a phone, a syllable, a word, a sentence, or some other desirable part of speech, depending on the application. Secondly, a context template, [$\mathcal{T}$; *R*; *s*, *t*], is generated for the focus unit. Thirdly, by matching against the template, an appropriate hypothesis set H([$\mathcal{T}$; *R*; *s*, *t*]) is determined. Depending on how stringent the template constraints are, the hypothesis set under examination may be greatly narrowed, over traditional GPP approaches. The Template Constrained Posterior (TCP) of [$\mathcal{T}$; *R*; *s*, *t*] is the generalized posterior probability summed on all the string hypotheses in H([$\mathcal{T}$; *R; s, t*]), as in Eq. 4.

$$P([\mathcal{T};R;s,t]\,|\,x_1^T)=\sum_{\substack{N,h=[w;s,t]_1^N\\ h\in H([\mathcal{T};R;s,t])}}\frac{\prod_{n=1}^{N}p^{\alpha}(x_{s_n}^{t_n}\,|\,w_n)\cdot p^{\beta}(w_n\,|\,w_1^N)}{p(x_1^T)} \qquad (4)$$

where x_1^T is the whole sequence of acoustic observations, α and β are the exponential weights for the acoustic and language model likelihoods, respectively. In calculating TCP, the reduced search space, the time relaxation registration, and the weighted acoustic and language model likelihood are handled similarly as in GPP [2]. The difference between the TCP and GPP is the determination of the string hypotheses set, which corresponds to the term under the sigma summation notation.

The posterior probability calculated can be utilized to identify potential errors between the audio recording and the transcription.

3.4. Advantages of TCP

The TCP approach examines both the focused unit and the context to the left and right of the focused unit. In this way, TCP better discriminates competing phones, because a string hypothesis with competing phones is less likely to match both the (partial) context and the actual (focus) phone. Therefore, hypotheses containing the competing phone will be less likely to be included in the TCP calculation, a significant advantage over the traditional GPP approaches [6].

Furthermore, the TCP approach provides additional robustness against incorrect time boundaries [6]. In the standard GPP approach, the focus unit is expected to appear within a narrow timeframe. TCP however supports a broader timeframe in the calculation, in order to examine the context. Thus, the TCP approach is more robust against incorrect time boundaries, which may be caused by insertion, deletion, or substitution errors.

Essentially, the proposed template constrained approach uses templates to limit the hypothesis set during the posterior probability calculation for a selected focus unit. These templates may be tailored to provide a fine degree of granularity, from specifically defined context to loosely defined contexts.

4. VERIFYING PHONE TRANSCRIPTION BY TCP

4.1. System flowchart

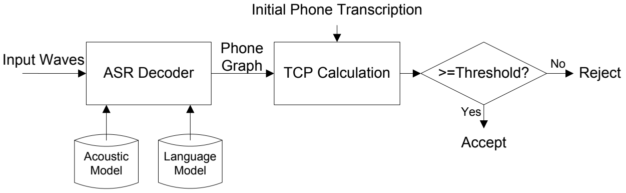

Figure 4: *A flowchart of verifying phone transcription by TCP.*

Phone level TCP is used as the confidence measure to identify potential phone errors in transcriptions. Figure 4 depicts the flowchart of verifying phone transcription using TCP. Firstly, with acoustic model and language model, ASR phone decoder generates phone graph for a spoken input. By regarding each phone in the initial phone transcription as a focused phone, its context template is generated (as described in Section 3.2) and its TCP is calculated. The calculated TCP scores are compared with a threshold, which is optimized upon a development set. A phone will be accepted if its TCP is higher than the threshold. Otherwise, the phone will be rejected. For practical use, the verification decision is made at sentence level. A sentence will be rejected if it contains one or multiple phone errors.

4.2. Experimental setup

We evaluate the proposed method on an American English TTS database, recorded by a female native speaker. It consists of nearly 20,000 utterances and has a rich coverage of different phonetic contexts. Four transcribers manually verified 1,234 sentences and pinpointed the phone errors or the disagreements between the audio recordings and the original transcriptions. The verified transcription serves as correct reference in later phone verification experiments. The phone errors are classified into two different types, major error and minor error. Minor error refers to the errors between confusable phone pairs, which are difficult to differentiate even for an experienced annotator. The minor errors include confusions between /ih/- /iy/, /uw/-/uh/, /ax/-/ah/, and etc. Major phone errors, on the other hand, are more serious and they can impair the synthesized voice quality appreciably. Among the 1,234 sentences, more than half of the sentences contain either major or minor errors. In average, every erroneous sentence contains 1.6 phone errors.

We split these 1,234 sentences into two, development and test, sets, each consisting of 617 sentences. All erroneous sentences are evenly distributed into the development and test sets, as shown in Table 1. The acoustic model is adopted from Microsoft Speech API, which is a gender-dependent, tri-phone HMM model trained by thousands of hours of speech data. A phone tri-gram model is served as the language model. For each spoken utterance, a dense phone graph is generated by a Viterbi decoder with a wide-beam. The phone graph density (GD), the graph error rate (GER), and phone error rates (PER) of the development set and test set are listed in Table 2.

Table 1: *Initial transcription for Development set and Test set.*

Data Set	#Sentence	#Error Sentence (Minor)	#Error Sentence (Major)	#phone error (Minor)	#phone error (Major)
Development	617	142	183	202	335
Test	617	105	220	161	367

Table 2: *Phone graph for Development set and Test set.*

Data Set	Phone Error Rate (PER%)	Graph Error Rate (GER%)	Graph Density (Edges/phone)
Development	77.78	97.19	531
Test	78.03	97.07	518

Acceptance ratio is defined as the ratio of accepted sentences to total sentences. Error hit rate in rejected sentence is the ratio of the hit erroneous sentence to all the rejected sentences. We assess the TCP verification performance by looking at the relation between acceptance ratio and error hit rate in rejected sentences.

4.3. Configuration optimization on the development set

To minimize phone verification errors on the development set, we searched exhaustively for the optimal configuration settings of the template generation and the TCP verification threshold, T. In template generation, context window length (2L+1), partial match ratio (R) and the KLD threshold (T_{KLD}) are the three parameters to be determined. Verification under different (2L+1, R, T_{KLD}, T) combinations has been tested on the development set, with the verification results plotted in Figure 5. Each point corresponds to the TCP performance under a unique configuration setting. The envelop line is of particular interests, which can be regards as the best performance of this method. We can see that, when the acceptance rate is 92%, 97% of the rejected sentences contain errors (refer to the envelop line of the red curves), and 90% of the rejected sentences contain major errors (refer to the envelop line of the green curves). When the acceptance ratio is 80%, nearly 80% of the rejected sentences contain errors, and about 60% of them contain major errors. The configuration settings along the envelop line, which can be regarded as the optimal configurations under different acceptance ratio, is then used in the following experiments on the test set.

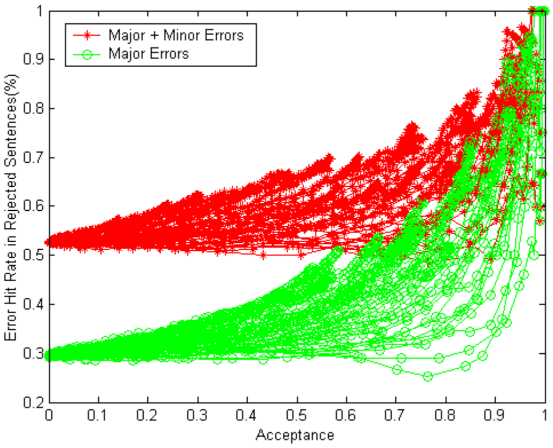

Figure 5: *TCP performance on the development set.*

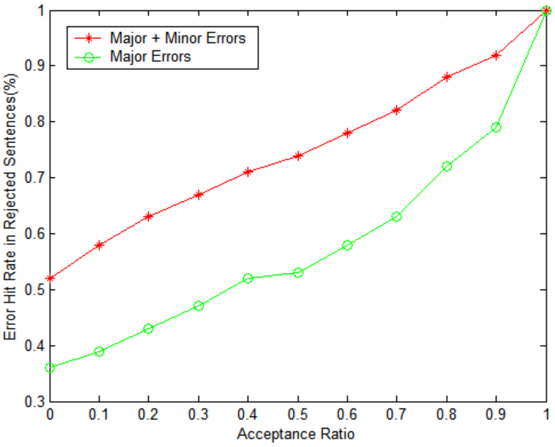

Figure 6: *TCP performance on the test set.*

4.4. TCP performance on the test set

With the obtained optimal configurations, we test the TCP performance on the test set. As shown in Figure 6, when the acceptance ratio is 90%, 92% of the rejected sentences contain errors, where 72% of them contain major errors. When the acceptance ratio is 80%, 88% of the rejected sentence contains errors, where 63% of them contain major errors. This result shows that phone level TCP is effective for detecting potential phone errors. The results also show that major errors can be detected more easily than the minor ones. The reason is that TCP is computed based on a phone graph, a byproduct of ASR decoding. Due to the inadequate discrimination of acoustic HMM, the models cannot adequately differentiate certain phonemes, similarly for TCP. This method can dramatically reduce manual work by directing human effort to the verification of the unconfident sentences.

5. CONCLUSIONS

A new confidence measure, TCP, is proposed for verifying transcription errors. Different from the standard GPP, the TCP approach examines both the focused unit and the context to the left and right of the focused unit. The concept of *template* is also proposed to limit the hypothesis set used in calculating the posterior probability for a selected focus unit. These templates may be flexibly tailored to provide various granularities, from a finely defined context to loosely defined contexts. We tested this method for detecting the potential errors in phone transcription, which is based on a template for a focused phone. The optimal configuration of TCP, in term of context window length, optimal partial match ratio, optimal KLD threshold, and decision threshold, is also studied in the paper. Evaluated on a corpus used by an English TTS system, which contains 32.6% sentence errors and 0.62% phone errors, TCP achieves error hit rate in rejected sentences of 92% and 88% respectively when the acceptance ratio is 90% and 80%. Future research will focus on applying TCP to speech recognition systems.

6. REFERENCE

[1] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Proc.,* Vol. 9, pp.288-298, 2001.

[2] F.K. Soong, W.K. Lo, and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," in *Proc. SWIM-2004, Hawaii, 2004.*

[3] D. Binnenpoorte, and C. Cucchiarini, “Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality,” in *Proc. ICPhS-2003*, 2003.

[4] T.J. Hazen, “Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings,” In *Proc. INTERSPEECH-2006*, pp. 1606-1609, Pittsburgh, Pennsylvania, September 2006.

[5] L.J. Wang, Y. Zhao, M. Chu, F.K. Soong, and Z.G. Cao, “Phonetic transcription verification with generalized posterior probability,” in *Proc. INTERSPEECH-2005*, Lisbon, 2005.

[6] H. Zhang, L.J. Wang, F.K. Soong, “Context Constrained-Generalized Posterior Probability for Verifying Phone Transcriptions,” in *Proc. INTERSPEECH-2007*, Antwerp, 2007.

[7] M. Saraclar, R. Sproat, “Lattice-Based Search for Spoken Utterance Retrieval,” in Proc. HLT’2004, Boston, 2004.

[8] P. Liu, F.K. Soong, J.L. Zhou, “Effective Estimation of Kullback-Leibler Divergence between Speech Models”, Microsoft Research Asia, Technical Report, 2005.