Context Constrained-Generalized Posterior Probability for Verifying Phone Transcriptions

Hua Zhang [1], *Lijuan Wang* [2], *Frank Soong* [2], *Wenju Liu* [1]

[1] Institute of Automation, Chinese Academy of Sciences
[2] Microsoft Research Asia, Beijing, China

{hzhang,wjliu}@nlpr.ia.ac.cn, {lijuanw, frankkps}@microsoft.com

Abstract

A new statistical confidence measure, Context Constrained-Generalized Posterior probability (CC-GPP), is proposed for verifying phone transcriptions in speech databases. Different from generalized posterior probability (GPP), CC-GPP is computed by considering string hypotheses that bear a focused phone with partially matched left and right contexts. Parameters used for CC-GPP include context window length, a minimal number of matched context phones, and verification thresholds. They are determined by minimizing verification errors in a development set. Evaluated on a test set of 500 sentences that consist of 2.1% phone errors, CC-GPP achieves 99.6% accuracy and 78.7% recall when 90% of the phones are accepted.

Index Terms: Context Constrained pattern, posterior probability, CC-GPP, confidence measure

1. Introduction

Large, well-annotated speech corpora have become almost indispensable for speech research and product/service development. Take the concatenation-based Text-to-Speech (TTS) synthesis as an example [1]. The quality of the synthesized speech depends on the accuracy of annotated phonetic labels and corresponding contexts for selecting good acoustic units from a pre-recorded corpus. However, annotation of a large corpus usually requires extensive manual work, which can be time-consuming, costly, and remains prone to human errors. Recourse to automatic or semiautomatic annotation of speech data is therefore desirable, such as forced recognition (forced alignment). These techniques are more efficient than human checking but annotation errors can still be made for the following reasons:

- Reading errors, or orthographic pronunciation errors, occur;
- Incomplete lists of all possible pronunciations of a word in the lexicon and letter-to-sound errors for out-of-vocabulary words may occur;
- Idiosyncratic pronunciations of individual speakers cause inaccuracies.

Confidence measures are useful for verifying speech transcriptions by assessing the reliability of a focused unit, e.g., word or phone. Various approaches for measuring confidence of speech transcription have been attempted [3][4][5] . They can be roughly classified into three categories: i) feature based; ii) explicit model based; and, iii) posterior probability based. Feature-based approaches try to assess the confidence based on selected features (e.g., word duration, part-of-speech, acoustic and language model back-off, word graph density, or the like) using trained classifiers. Explicit model-based approaches employ a candidate class model with competing models (e.g., an anti-model or a filler model) and a likelihood ratio test is applied. The posterior probability-based approach tries to estimate the posterior probabilities of a recognized entity given all acoustic observations [2].

In this study we propose a new confidence measure, *Context Constrained-Generalized Posterior Probability* (CC-GPP), for verifying phone transcriptions in speech databases. A flexible but cogent, context constrained pattern, is incorporated in computing CC-GPP. Tested on an English TTS corpus, we show that confidence and robustness against incorrect phone boundaries are improved as compared to the standard GPP.

The rest of the paper is organized as follows. In section 2 we review briefly the Generalized Posterior Probability (GPP). In section 3, CC-GPP is proposed. In section 4, CC-GPP is used in verifying phone transcriptions. In section 5 we present our experimental results. In section 6 a conclusions are drawn.

2. Generalized Posterior Probability

Generalized posterior probability (GPP) is a probabilistic confidence measure for verifying recognized (hypothesized) entities at a subword, word or string level [2]. It was applied to utterance verification under various testing conditions, e.g., [6][7]. GPP at word level assesses the reliability of a focused word by "counting" its reappearances in the word graph and weighting the corresponding acoustic and language model likelihoods exponentially and then normalizing the weighted reappearances by the total acoustic probability. Word level GPP is defined as

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

where the triple [*w*; *s, t*] is the focused word *w* with its starting time *s* and ending time *t*; x_1^T is the whole sequence of acoustic observations; *N* is the number of words of a string in the graph; α and β are the exponential weights for the acoustic and language model likelihoods, respectively. $P(x_1^T)$, the acoustic probability of all observations, can be computed by summing the likelihoods that are similarly weighted as in the numerator, of all paths in a word graph.

3. Context Constrained-Generalized Posterior Probability (CC-GPP)

3.1. Overview

A string hypothesis is an ordered sequence of recognized entities, e.g., phones or words. 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 like a decoded word graph. H, a subset of R, contains all string hypotheses that contain the focused word *"w"* with a given time range of starting and ending points. The posterior probability of *"w"* can be obtained in Equation (2), i.e., the quotient of 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 posterior probability $P(w|x_1^T)$ for verification.

$$p(w \mid x_1^T) = \frac{\sum_{h \in H} p(h)}{\sum_{h \in R} p(h)}, \quad H \subset R \tag{2}$$

In standard GPP computation, the correct hypotheses set H for [*w*; s, t], defined in equation (1), is obtained by finding every string hypothesis that contains the focused entity *"w"* and intersects with the specified time interval [s, t]. This method selects the correct hypothesis only based on the focused entity, which may cause problems when the start/end time of the focused entity is incorrect or the reappearance of a competing word grows up when the graph gets rich.

To avoid these problems, we propose a *context constrained pattern* as a template to sift out the string hypotheses that carry information of the decoded entity and its contexts for H. Different from the standard GPP, the context constrained pattern selects a string hypothesis based upon not only the focused entity, but also the partially matched, left and right contexts. The advantages are: 1) the tolerance to time boundary variation is improved when the string hypothesis is evaluated on a wider context window, hence its time range; 2) confidence is enhanced because a string hypothesis bearing the focused entity along with its neighboring, partially matched contexts is more reliable than that with the focused entity alone. The GPP calculated upon H thus becomes the *Context Constrained Generalized Posterior Probability* (CC-GPP).

3.2. Context constrained pattern

Here we use *word* as the decoded entity to demonstrate the CC-GPP approach, which is readily applicable to other entity levels, such as phone, or subword.

Let $\sum$ be a finite word *vocabulary*, and let $\sum^*$ be the set of all *strings* over $\sum$. Denote $|v|$ the length of a string *v*. String $v = v_1 \cdots v_p \in \sum^*$ is called a substring of string $h = h_1 \cdots h_n \in \sum^*$ if there exists $1 \le i \le (n-p+1)$ such that $v_j = h_{i+j-1}$ for $1 \le j \le p$. e.g.: "$\underline{w_3 w_4 w_5}$" is a substring of "$w_1 w_6 \underline{w_3 w_4 w_5} w_8 w_7$".

Definition of Context Constrained Pattern: A context constrained pattern is a triple $[w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m]$, where $w_k \in \sum$ is the focused word, $w_{k-L} \cdots w_k \cdots w_{k+L} \in \sum^*$ is a word string covering the *L* context words to its left and right, respectively, m is a non-negative integer ($m \le 2L$), the minimal matched word number among the *2L* context words. A context constraint pattern $[w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m]$ matches a string hypothesis $h \in \sum^*$ if a substring v of h can be made from $w_{k-L} \cdots w_k \cdots w_{k+L}$ with (*2L-m*) or less substitutions among the *2L* context words of w_k. In other words, more than (*m+1*) words, including the w_k, in the substring v are the same as the corresponding (*m+1*) words in $w_{k-L} \cdots w_k \cdots w_{k+L}$.

To illustrate this we use Figure 1, a string hypothesis that matches $[w_k; w_{k-3} w_{k-2} w_{k-1} w_k w_{k+1} w_{k+2} w_{k+3}; 3]$ satisfies two conditions: 1) the focused word w_k is matched; 2) at least three context words, which can be either left or right contexts , need to be matched at the correct context positions.

For example: $[w_4; \underline{w_1} w_2 \underline{w_3} w_4 w_5 w_6 \underline{w_7}; 3]$ matches "$\underline{w_1} w_6 \underline{w_3} w_4 w_9 w_8 \underline{w_7}$", but does not match "$\underline{w_1} w_3 w_2\ w_4 \underline{w_5} w_8 w_6$", whereas, $[w_4; \underline{w_1} w_2 w_3 w_4 \underline{w_5} \underline{w_8} w_7; 3]$ will match "$\underline{w_1} w_3 w_2 w_4 \underline{w_5} \underline{w_8} w_6$".

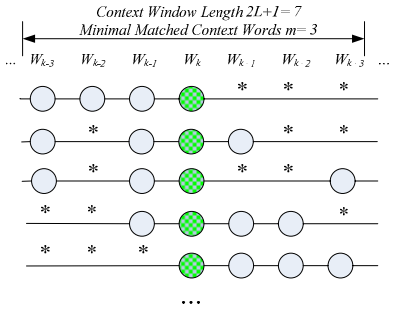

Figure 1: *Illustration of context constrained pattern, where * is a "don't care" symbol, which can match any word.*

The context constrained pattern incorporates the contextual information in a constrained but still flexible way. On the one hand, altogether *2L* context words before and after the focused center word are considered; on the other hand, only partial, not fully matched *2L* context words are required. The matching template is intentionally designed with some "don't cares" to relax the context constraint.

3.3. Context Constrained GPP

All string hypotheses that match $[w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m]$ forms the hypothesis set H, denoted as $H([w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m])$. The Context Constrained-Generalized Posterior Probability (CC-GPP) of w_k is the generalized posterior probability sum of all the string hypotheses in $H([w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m])$.

$$P([w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m] \mid x_1^T) = \sum_{\substack{N, [s_1^N] \\ h \in H([w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m])}} \frac{\prod_{n=1}^{N} p^{\alpha}(x_{s_n}^{t_n} \mid w_n) \cdot p^{\beta}(w_n \mid w_1^N)}{p(x_1^T)} \tag{3}$$

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.

The Context Constrained GPP calculation is similar to that in [6]. The reduced search space, the time relaxation registration, and the weighted acoustic and language model likelihood are handled as follows.

- A decoded hypotheses graph is served as the reduced search space.
- It is desirable to relax the time registrations in finding out all matched string hypotheses. If a substring of *h* that matches $[w_k; w_{k-L} \cdots w_k \cdots w_{k+L}; m]$ exists and intersects the time interval $[s_{k-L}, t_{k+L}]$, the string hypothesis is included in CC-GPP calculation.
- The acoustic and language model likelihood weightings are adjusted to prevent the posterior probability from being dominated by just a few high likelihood strings, and to accommodate the modeling discrepancies in the practical implementations.

3.4. Advantages of CC-GPP

The main idea in the proposed CC-GPP is that a context constrained pattern template is used to select appropriate

string hypotheses. Not only the focused entity, but the partially matched contexts to its left and right are considered. The advantages of CC-GPP are on two aspects: discrimination against competing words and robustness against incorrect given time boundaries.

3.4.1. Discrimination against competing words

Competing words in a graph usually weaken the confidence of GPP, especially for a large graph. A good confidence measure should suppress the competing words to improve correct verification of the true words. A string hypothesis bearing a competing word has less possibility of containing partially matched context than it does of bearing the true word. So, it is more possible that a context constrained pattern template will sift out string hypotheses that bear the true word as a center entity, while rejecting those hypotheses bearing the competing word as the center. As a result, hypotheses bearing a competing word will be less likely to be included in CC-GPP calculation. Therefore, the confidence of CC-GPP is enhanced over that of GPP.

3.4.2. Robustness against incorrect boundaries

Given a speech transcription containing substitution, insertion, and deletion errors, it is inevitable that incorrect phone boundary exists. So, making the confidence measure to be robust against incorrect time boundaries is important. Using CC-GPP, the desired robustness is improved when evaluating a string hypothesis on a wider context, and hence a larger time range. In standard GPP, when a focused word $[w_k; s_k, t_k]$ is given, the time registration is relaxed to overlap with $[s_k, t_k]$. However, in CC-GPP, the time registration is further relaxed to overlap with $[s_{k-L}, t_{k+L}]$, where $2L+1$ is the context window length in the CC-GPP calculation. Therefore, the time range in CC-GPP calculation is much wider than that in GPP.

4. Phonetic Transcription Verification with Context Constrained-GPP

The proposed CC-GPP is a general confidence measure for verifying the hypothesized entities at phone, syllable or word, levels. Here, we test CC-GPP in phone transcription verification.

4.1. Phone verification with CC-GPP

Phone level CC-GPP is used as the confidence measure to assess the reliability of each phone in transcriptions. The search space is reduced to a phone graph. As defined in 3.2, all phone string hypotheses that match $[p_k; p_{k-L} \cdots p_k \cdots p_{k+L}; m]$ are added to the hypothesis set H, i.e., $H([p_k; p_{k-L} \cdots p_k \cdots p_{k+L}; m])$. CC-GPP of p_k can thus be computed by summing up GPP's of all string hypotheses in $H([p_k; p_{k-L} \cdots p_k \cdots p_{k+L}; m])$. To minimize phone verification errors, the optimal context window length ($2L+1$), the minimal number of matched context phone m, and the decision threshold T are optimized on a development set. With the optimized configuration, the phone level CC-GPP is calculated. A phone is accepted when its CC-GPP is higher than the threshold T; otherwise, it will be rejected.

4.2. Data preparation

The speech corpus for our experiments is an English database, recorded by a professional male speaker for constructing a TTS system. It consists of 4,273 read utterances and various phonetic contexts are covered.

There are two manual transcriptions of the corpus, initial transcription with 10.5% sentence error rate and 0.65% phone error rate, and a proofread transcription which is manually verified by several transcribers. The proofread transcription is served as a correct reference during the phone verification.

In particular, a development set and a test set, each consisting of 500 sentences, are used to evaluate CC-GPP. All errors are evenly distributed into the development set and the test set, as shown in Table 1. Other than the initial transcription and its proofread version, an artificially created transcription with substitution errors is generated and used for optimizing the parameters of the context constrained pattern. The artificial transcription is generated from the proofread transcription, by substituting one phone at a time in each sentence with all other phones in the set.

4.3. Phone graph generation

The whole corpus, with its initial transcription, is used to train the speaker dependent acoustic HMMs. We chose 39 acoustic features (12MFCC+12ΔMFCC+12ΔΔMFCC+logE+ΔlogE+ΔΔlogE). Four Gaussian components per mixture are used for modeling the output probability density function of each state of a tied tri-phone. A position dependent phone bigram language model is used to generate a rather dense phone graphs with a wide-beam in Viterbi search. The phone graph density (GD) and graph error rate (GER) and string and phone error rates of two initial and artificial, transcriptions, are listed in Table 1.

Table 1: *Development set and test set.*

Data Set	Phone number	Phone Graph		Initial Transcription		Artificial Transcription	
		GD	GER%	SER%	PER%	SER%	PER%
Development Set	25,794	9,450	1.73	44.8	2.12	100	1.51
Test Set	26,745	9,321	1.66	45.0	1.91	--	--

4.4. Evaluation method and Baseline

Given a decision threshold, each phone in transcription is either accepted or rejected according to its CC-GPP. False Acceptance Rate (FAR) and False Rejection Rate (FRR) are used to evaluate CC-GPP performance. Equal Error Rate (EER) is defined as the point where FAR and FRR are equal.

Acceptance ratio is defined as the ratio of accepted phones to total phones. Recall is the percentage of rejected error phones in the total error phones. Accuracy is the percentage of correct findings in accepted phones.

Phone verification using standard phone level GPP is applied to the baseline system described in section 2.

5. Experimental Results

In the first experiment, we study various parameters for the context constrained pattern on the development set. Then, we compare CC-GPP with GPP on the robustness on incorrect phone boundaries. In the third experiment, phonetic verification using CC-GPP is evaluated on the test set.

5.1. Optimal context window length and minimal number of matched context phones

In a context constrained pattern, appropriate context window length $2L+1$ and minimal number of matched context phones m are two parameters to be determined. CC-GPP configured with different ($2L+1$, m) combinations was evaluated on the

development set using artificially created substitution errors. Under different (*2L+1, m*) settings, the Equal Error Rate (EER) is depicted in Figure 2. It shows that a context window length *2L+1=7* and minimal number of matched context phones *m=3* yield a good EER point. This configuration is then used in the following experiments.

We assessed the confusion between each phone pair (p_a, p_b) defined in Equation (4). In the top 10 confusable phone pairs, which are not symmetrical, as presented in Table 2, the confusion rate of all phone pairs is consistently reduced, by 7%~37%, relatively. This result confirms the CC-GPP capability for discriminating competing phones.

$$Confusion(p_a; p_b) = \frac{\text{No. of accepted } p_b \text{ that substituted } p_a}{\text{No. of all } p_b \text{ that substituted } p_a} \quad (4)$$

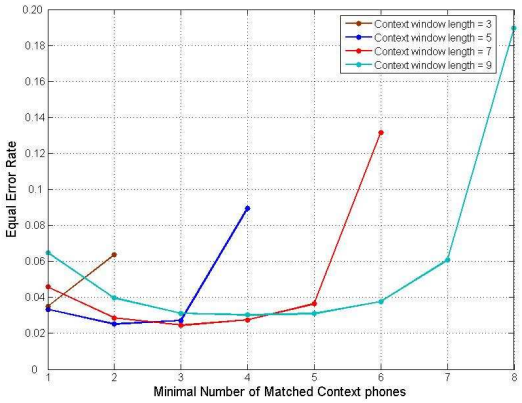

Figure 2: *EER of CC-GPP with different context window length and minimal matched phone number.*

Table 2: *Improved Discrimination of CC-GPP over GPP in confusable phone pairs.*

Confusion rate	Top 10 confusable phone pairs									
	ch;t	ch;sh	th;s	aa;ah	th;t	ae;eh	uh;ax	m;n	s;z	d;t
GPP	0.63	0.54	0.49	0.49	0.42	0.41	0.38	0.43	0.34	0.37
CC-GPP	0.57	0.34	0.38	0.41	0.39	0.36	0.25	0.23	0.22	0.31
Relative reduction	10%	37%	22%	16%	7%	12%	34%	32%	35%	16%

5.2. Robustness against Incorrect Boundaries

This experiment is designed to test the robustness of CC-GPP when the given phone boundaries are perturbed. In the development set, the correct phone boundaries are perturbed by a specified value and corresponding EER is evaluated. As shown in Figure 3, EER of GPP increases rapidly as the boundary is perturbed by more than 30ms, while CC-GPP maintains its high performance over a broad range of perturbation (<300ms). The results show that with a long, constrained context window and the requirement of matched context positions, the proposed CC-GPP is more relaxed in its boundary precision requirement than the standard GPP.

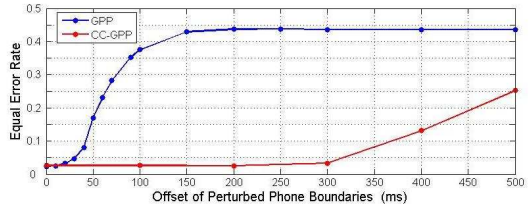

Figure 3: *EER of CC-GPP and GPP given perturbed incorrect boundary.*

5.3. Performance on test set

We performed phone verification experiments by using GPP and CC-GPP on the test set. The EER of CC-GPP is 15.3%, which is better than that of GPP 18.1%, by 15.5%. Recall and accuracy at different acceptance ratios are presented in Figure 4. Using CC-GPP, when 90% of data are accepted, the recall and accuracy are 78.7% and 99.6%, respectively.

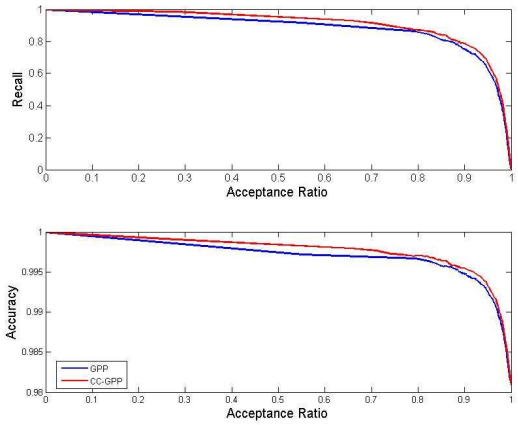

Figure 4: *Recall and accuracy of CC-GPP and GPP at different acceptance ratio.*

6. Conclusions

A new confidence measure, CC-CPP, is proposed for verifying phone errors in transcription. Different from the standard GPP where the string hypotheses bearing reappearance of the focused center phone are used, an additional constraint of partially matched neighboring contexts is imposed. Context window length, minimal matched phones, and decision threshold are determined by minimizing verification errors in a development set. Evaluated on a test set from an English TTS corpus, which contains 2.1% phone errors, 99.6% and 78.7% of accuracy and recall are obtained, respectively, when 90% phones are accepted. These findings indicate that using CC-GPP for verifying phone transcriptions in speech databases can provide superior confidence and robustness against incorrect phone boundaries as compared to the standard GPP. Next, we will use CC-GPP to bootstrap speech recognition systems.

7. References

[1] Fackrell, J., Skut, W., and Hammervold, K. "Improving the accuracy of pronunciation prediction for unit selection TTS," in *Proc. EUROSPEECH-2003, pp. 2473-2476, Geneva, Switzerland, September 2003*.

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

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

[4] Jordi, A., Pablo D. A., and Antonio B., "Database Pruning for unsupervised building of Text-to-Speech voices," In *Proc. ICASSP-2006*, pp. 889-892, Toulouse, France, May 2006.

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

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

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