TONE-ENHANCED GENERALIZED CHARACTER POSTERIOR PROBABILITY (GCPP) FOR CANTONESE LVCSR

Yao Qian[1,2] *Frank K. Soong*[1,2] *Tan Lee*[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2]Microsoft Research Asia, Beijing, China
{yaoqian, frankkps}@microsoft.com, tanlee@ee.cuhk.edu.hk

ABSTRACT

Tone-enhanced, generalized character posterior probability (GCPP), a generalized form of posterior probability at subword (Chinese character) level, is proposed as a rescoring metric for improving Cantonese LVCSR performance. The search network is constructed first by converting the original word graph to a restructured word graph, then a character graph and finally, a character confusion network (CCN). Based upon GCPP enhanced with tone information, the character error rate (CER) is minimized or the GCPP product is maximized over a chosen graph. Experimental results show that the tone enhanced GCPP can improve character error rate by up to 15.1%, relatively.

1. INTRODUCTION

Most HMM based speech recognizers search for the word string (sentence) hypothesis that yields the maximum a posterior (MAP) probability. Under the MAP criterion misrecognized sentences are minimized in the expected value sense. However, word error rate (WER), rather than sentence error rate, is more universally accepted in the speech recognition community as the sole objective performance measure of an LVCSR system. Many studies have been done on how to train a recognizer or perform search in recognition to optimize such measure. For example, minimizing the expected word error rate was proposed as the search criterion for speech recognition [1-4]. Estimation of word posterior probability and determination of the sentence with minimum expected word error were investigated for N-best output [1]. They were also applied to a word graph [2], where multiple string alignment instead of pairwise string alignment was adopted. In [4], the minimum Bayes-risk (MBR) approach, a more general cost function based on word error measurement, is implemented to rescore N-best list and to A* search over the word lattice. In addition, confidence measures at the word level were used for rescoring [5-7].

Cantonese, a popular Southern Chinese dialect, is a syllabically paced, tonal language of which tones are lexical. The basic written unit of Cantonese is the Chinese character. Each character is pronounced as a tonalized monosyllable, which has a relatively simple (C)-V-(C) structure and relatively stable duration than other speech units in Chinese. Character, a subword unit in Chinese, also plays an important role in both morphology and phonology of Chinese languages. Most of the morphemes consist of one single character. In written Chinese, except for the occasional punctuation marks, there is no delimiter (like blank space) between two adjacent characters. As a result, the definition of a word in Chinese is somewhat vague and the performance of Chinese LVCSR is usually measured by the corresponding character error rate (CER), rather than the word error rate. In this paper, we propose to use tone-enhanced, generalized character posterior probability (GCPP) as a rescoring metric for Cantonese LVCSR. GCPP is computed in a restructured word graph by incorporating the tone information. Two improved search approaches based on GCPP, either minimizing character error rate (CER) or maximizing GCPP product, will be presented.

2. GENERALIZED CHARACTER POSTERIOR PROBABILITY (GCPP)

Posterior probability assesses quantitatively the correctness of recognition results. It can be computed at sentence, word or subword, e.g., syllable or character, level. There have been numerous studies on its estimation and applications [8][9]. Generalized posterior probability [10] tries to address the various modeling discrepancies and numerical issues in computing the posterior probability. It is designed to incorporate automatically trained optimal weights to equalize the different dynamic range of acoustic and language models, segmentation ambiguities, etc. It attempts to configure the most appropriate posterior probabilities for different recognition or verification tasks. Its effectiveness has been demonstrated in verification of recognition outputs under both clean and noisy conditions [11][12].

2.1 GCPP estimation

The posterior probability of a specific character can be estimated by summing up the posterior probabilities of all string hypotheses that contain the same character with identical starting and ending time. The output of a Chinese recognizer is a string of words, $w_1^M = w_1, w_2, ..., w_M$. It can be further decomposed into a sequence of characters, $c_1^L = c_1, c_2, ..., c_L$. GCPP computation for a specific character c can be defined as,

$$p([c;s,t] \mid x_1^T) = \sum_{\substack{L,[c';s',t']_1^L \\ \exists n, 1 \le n \le L \\ [c';s',t']=[c_n;s_n,t_n]}} \frac{\prod_{l=1}^{L} p^{\alpha}(x_{s_l}^{t_l} \mid c_l) p^{\beta}(c_l \mid c_1^{l-1})}{p(x_1^T)} \qquad (1)$$

where $p(x_{s_l}^{t_l} \mid c_l)$ is the acoustic model (AM) likelihood of character c_l with starting time s_l and ending time t_l; $p(c_l \mid c_1^{l-1})$

is the prior probability of character c_l, or the language model (LM) likelihood, given all its history c_1^{l-1}; and α and β are the exponential AM and LM weights that are jointly optimized with a held-out set of data. The character acoustic likelihood and boundary information can be recorded during the first-pass Viterbi search, but the character prior probability is not available when word-based language model is used in our Cantonese LVCSR, which will be introduced in Section 4.1. Character is a subword unit, the word w_m comprises $Z(\geq 1)$ such characters $w_m = c_{m,1}, c_{m,2}, ..., c_{m,Z}$. To compute GCPP, Eq. (1) is revised as

$$p([c;s,t] \mid x_1^T) = \sum_{\substack{M,[w;s,t]_1^M \\ \exists n, 1 \le n \le M; \exists j, 1 \le j \le Z; \\ [c;s,t]=[c_{n,j};s_{n,j},t_{n,j}]}} \frac{\prod_{m=1}^{M}\left[\prod_{z=1}^{Z} p^{\alpha}(x_{s_{m,z}}^{t_{m,z}} \mid c_{m,z}, w_m)\right] p^{\beta}(w_m \mid w_1^{m-1})}{p(x_1^T)} \quad (2)$$

GCPP [18] can be estimated from generalized word posterior probability (GWPP). A word and its constituent characters share the same string hypothesis. However, if the GCPP is simply made equal to the GWPP of its carrier word, it would be under-estimated. This is because the same character may appear in different words. If the word posterior probabilities are known, the GCPP $p([c;s,t] \mid x_1^T)$ could be estimated by summing up the posterior probabilities of all words containing this character over the same time interval, i.e.

$$p([c;s,t] \mid x_1^T) = \sum_{\substack{[w_m;s_m,t_m] \\ \exists z, 1 \le z \le Z \\ [c;s,t]=[c_{m,z};s_{m,z},t_{m,z}]}} p([w_m;s_m,t_m] \mid x_1^T) \quad (3)$$

where $p([w;s,t] \mid x_1^T)$ is the GWPP estimated from a restructured word graph by the forward-backward algorithm. With the character boundary information, the word graph can be converted into a character graph. For each character arc, the GCPP is initially assigned the GWPP value of the carrier word. If two arcs of character graph with the same character identity over the same time interval are merged into one arc, the GCPP of the resultant character arc should be the summation of those two arc scores. An example is illustrated in Fig. 1.

2.2 Restructured word graph

Different algorithms can be used to generate word graphs and the structure of a word graph is important to GCPP estimation. Our word graph is generated based upon word dependent lexical tree search with word pair constraint [13]. In the first-pass Viterbi search, word boundaries are optimized with given predecessor words and word acoustic scores are recorded at word-ending states. Word arcs with identical timing information are merged into a single node. The illustration of a word graph is shown in Fig. 2 (a), where arcs are labeled with their word identities, w_j, and the preceding words, v_i. Here the word score and timing information are not shown for clarity. Each arc is marked with the current and preceding word information, denoted as $w_j \mid v_i$. There are totally five legal paths from the start node to the end node as shown in Fig. 2 (a).

When estimating the generalized posterior probabilities, we restructure the word graph by ignoring all preceding word information, as shown in Fig. 2 (b). That is, a word pair is considered legal if two words are connected at a node and all connected paths in the resultant word graph are also legal. The word graph constructed in this way, due to its over-generation property, can recover promising hypotheses which might have been prematurely pruned in the first-pass. For example, the total number of legal paths in Fig. 2 (b) is increased from five to seven. A similar strategy was employed in [9].

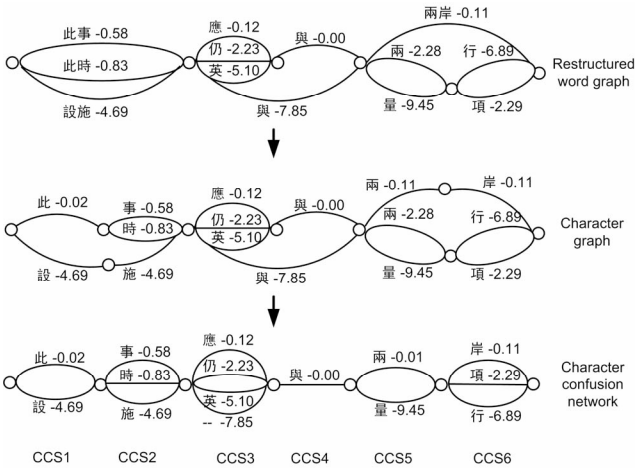

Fig. 1 An example of transforming word graph into character confusion network, each arc is associated with character and its logarithm of generalized posterior probability. The deletion in confusion set is denoted by "--".

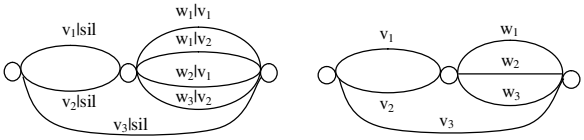

Fig. 2 The illustration of (a) a word graph and (b) restructured word graph

2.3 Enhancing GCPP by tone model

In Chinese, each character is pronounced as a tonalized monosyllable. The features derived from syllable-wide F0 contours in Chinese were shown to be effective for tone modeling and recognition. Our Cantonese tone model which characterizes not only the tone contour of a single syllable but those of the adjacent ones significantly outperforms previously reported methods [14]. We propose to use this tone model (TM) to calculate GCPP together with AM and LM. Eq. (1) is rewritten as

$$p([c;s,t] \mid x_1^T) = \sum_{\substack{L,[c;s,t]_1^L \\ \exists n, 1 \le n \le L \\ [c;s,t]=[c_n;s_n,t_n]}} \frac{\prod_{l=1}^{L} p^{\alpha}(x_{s_l}^{t_l} \mid qs_l) p^{\beta}(c_l \mid c_1^{l-1}) p^{\gamma}(y_{s_l}^{t_l} \mid qt_l) p(qt_l \mid c_l)}{p(x_1^T)} \quad (4)$$

where x and y denote the spectral features (e.g., MFCC) and pitch-related features such as F0, respectively. q_l is the tonal syllabic transcription of character c_l. It is composed by the syllabic transcription qs_l and the tonal transcription qt_l. γ is the weight of TM and should be adjusted to optimize the CER together with the weights of AM and LM.

3. GCPP-BASED RESCORING

GCPP provides a quantitative estimate for the correctness of recognized characters. It is more appropriate as a performance metric since the performance of Chinese LVCSR is usually measured by CER. Here, two improved search criteria based on GCPP are investigated.

3.1 Minimum character error rate

Minimum character error rate (MCER) search is similar to minimum word error rate search [2] but at subword level. In Section 2, we have converted the word graph into a character graph and computed GCPP for each character arc of the graph. Here, we construct character confusion network (CCN) based on character graph in order to implement the MCER search.

An arc clustering procedure [2], which is used to construct a word graph into a linear one, is modified to transform graph at character level. The clustering is performed in three stages: (1) pruning the arcs with low GCPP to avoid confusable clustering; (2) merging same character arcs with overlapping time intervals and assigning the summation of their GCPPs to the resultant arc; (3) grouping different character arcs into confusion sets according their time overlap, phonetic similarity and GCPP. The grouped character arcs are called the character confusion sets (CCS), which contain competing alternative character hypotheses with corresponding GCPPs. A sequence of CCSs form a linear graph called character confusion network (CCN). An example of converting word graph into CCN is shown as in Fig. 1. It consists of two transformations: (1) from the word graph to character graph and (2) from character graph to CCN.

In a CCN, each arc is labeled with a GCPP. MCER search can be expressed as

$$w^{*M}_1 = c^{*L}_1 = \arg\min_{\substack{c_1,\ldots,c_L \\ c_l \in cs_l}} \sum_l \sum_{\substack{c_j \in cs_l \\ c_l \neq c_j}} p(c_j \mid x_1^T) \qquad (5)$$

where cs_l denotes the l-th CCS, c_l and c_j are arbitrary character arcs in that CCS, and the word string w^{*M}_1 is equivalent to the character string c^{*L}_1. Eq. (5) can be achieved by selecting the character with the highest GCPP in each CCS. The sentence hypothesis with the lowest character error rate can be found by concatenating these characters.

3.2 Maximum GCPP product

Assuming no context dependencies, we can approximate sentence posterior by multiplying the posterior probabilities of all constituent characters. For ASR, the context dependencies of both acoustic observations and word in specific language should be considered. However, it is usually assumed that all observation frames are dependent only on the state that generates them, not on neighboring observation frames in conventional HMM based LVCSR. Moreover, a trigram language model used by forward-backward algorithm in computing character posteriors has taken the language context into account implicitly. Therefore, speech recognition can be viewed as finding a string which maximizes the product of GCPPs of individual characters, i.e.

$$\begin{aligned} c^{*L}_1 &= \operatorname*{argmax}_{L,[c;s,t]_1^L} p(c_1^L \mid x_1^T) \\ &= \operatorname*{argmax}_{L,[c;s,t]_1^L} \prod_{l=1}^{L} p([c_l;s_l,t_l] \mid x_1^T) \\ &= \operatorname*{argmax}_{L,[c;s,t]_1^L} \prod_{l=1}^{L} GCPP(c_l) \end{aligned} \qquad (6)$$

The graph-based DP search is applied to find 1-best path through the character graph.

4. EXPERIMENTS AND RESULTS

4.1 Speech database and baseline system

The speech corpus used in the experiments is CUSent™, which was collected at the DSP & Speech Technology Laboratory of the Chinese University of Hong Kong (CUHK) [15]. It is a continuous Cantonese speech corpus. The contents are given as in Table 1.

The baseline LVCSR system, named CURec, was also developed by the same research group at CUHK [16]. It uses context-dependent syllable Initial/Final models. The acoustic feature vector consists of 12 MFCC, log energy, and their first and second-order time derivatives. The Initial model is an HMM with 3 emitting states, while the Final model has either 3 or 5 emitting states, depending on its phonetic complexity. The output probability density function (pdf) of each emitting state was trained as a mixture of 16 Gaussian components. CURec uses a language model with 6,400 words. It was trained on a text corpus of 98 million Chinese characters from five Hong Kong newspapers.

A two-pass search algorithm is implemented in CURec. A word graph is generated in the first-pass search which is done time-synchronously with a word-conditioned lexical tree and a bigram LM. The second pass performs rescoring on the word graph with a trigram LM.

Table 1. Speech database used in this study

	Num of Sentences	Num of speakers	
		male	female
Training	20,378	34	34
Development	399	2	2
Testing	799	4	4

4.2 Performance of restructured word graph

Rescoring is performed in word graph (WG) and restructured word graph (RWG) with a trigram LM and re-optimized LM weight. The CER of LVCSR obtained from RWG outperforms that of WG by 1.56-2.03% absolute, as shown in Table 2. This confirms our conjecture that restructured word graph can recover some good paths pruned prematurely.

Table 3 gives the details of the RWGs used in our experiments. They are generated by the first-pass search with different beam widths. The recognition performance in terms of CER is also given. It is obtained by rescoring the RWG with a trigram LM. The graph error rate (GER) is computed by aligning the correct character sequence with the generated word graph to find the path with the least number of character errors. GER indicates the lower bound of the CER that is attainable by word graph rescoring. The word graph density (WGD) is the total number of word arcs divided by

the number of characters. In Table 3, RWG with the WGD of 21 gives almost the same CER and GER as the one with the WGD of 134. This suggests that an extremely wide beam-width may not be necessary for post-processing.

4.3 Results of GCPP-based rescoring

The CERs of recognition by using minimum CER and maximum GCPP product are shown at the bottom half of Table 2. Compared with the results of MAP, both approaches based on GCPP can reduce the absolute CERs by 0.32-0.74% at different beam widths. The performance of GCPP product is slightly worse than MCER. Table 2 also shows the performance of tone-enhanced GCPP based rescoring. GCPP enhanced with tone information results in absolute CER reductions of 2.29-2.80%, or 13.8-15.1% relative reduction. The greatest relative improvement of 15.1% is attained by MCER search for the wide beam generated graphs.

Table 2. Recognition performance in character error rate (CER)

		CER of recognition (%)		
Beam widths		Narrow	Medium	Wide
MAP (WG, trigram)		20.46	18.21	17.73
MAP (RWG, trigram)		18.58	16.18	16.17
GCPP	MCER	17.84	15.82	15.82
	GCPP product	18.03	15.86	15.82
Tone-enhanced GCPP	MCER	15.78	13.86	13.73
	GCPP product	16.01	13.89	13.85

Table 3. RWGs generated with using different beam widths

Beam width	CER (%)	GER(%)	WGD
Narrow	18.58	7.45	11
Medium	16.18	5.27	16
Wide	16.17	4.69	21
Widest	15.90	4.58	134

4.4 Optimal weights

Optimal AM, LM and TM weights in the above experiments are found from the development set. We adopt an efficient Downhill Simplex method [17] to perform the optimal weight search. The weights for graphs generated from different beam widths are trained independently. For example, the optimal weights for AM, LM and TM obtained in the case of medium beam width are $\alpha = 0.03$, $\beta = 0.9$ and $\gamma = 0.13$, respectively, and then are rather stable for different beam widths.

5. CONCLUSIONS

GCPP is proposed to be used as a search metric for improving Cantonese LVCSR performance. For each hypothesized character, Tone-enhanced GCPP is computed by incorporating the tone model score along with the corresponding acoustic and language model scores in a restructured word graph, which not only contains more string hypotheses than a typical N-best list but also recovers good but prematurely pruned string hypotheses. It is shown that in our two GCPP-based rescoring can reduce CER of recognition by 13.8-15.1% relatively at different beam width generated graphs.

6. REFERENCES

[1] Stolcke, A., Konig, Y. and Weintraub, M., "Explicit Word Error Minimization in N-best List Rescoring", *Proc. Eurospeech*, pp. 163-166, 1997.

[2] Mangu, L., Brill, E. and Stolckes, A., "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks", *Computer Speech and Language*, Vol.14, No.4, pp.373-400, 2000.

[3] Evermann, G. and Woodland, P. C., "Posterior Probability Decoding, Confidence Estimation and System Combination", *Proc. Speech Transcription Workshop*, 2000.

[4] Goel, V. and Byrne, W. J., "Minmimum Bayes-risk Automatic Speech Recognition", *Computer Speech and Language,* Vol. 14, pp.115-135, 2000.

[5] Wessel, F., Schluter, R. and Ney, H., "Using Posterior Probabilities for Improved Speech Recognition", *Proc, ICASSP*, 2000.

[6] Fetter, P., Dandurand, F. and Brietzmann, P. R., "Word Graph Rescoring Using Confidence Measures", *Proc. ICSLP*, 1996.

[7] Neti, C., Roukos, S and Eide, E., "Word-based Confidence Measures as a Guide for Stack Search in Speech Recognition", *Proc. ICASSP,* 1997.

[8] Weintraub, M. "LVCSR Log-likelihood Ratio Scoring for Key-word Spotting", *Proc. ICSLP*, 1995

[9] Wessel, F., Schluter, R., Macherey, K. and Ney, H., "Confidence Measures for Large Vocabulary Continuous Speech Recognition", IEEE Trans. On Speech and Audio Processing., Vol.9, No.3, pp.288-298, 2001

[10] Soong, F. K., Lo W. K. and Nakamura, S., "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words", *Proc. SWIM,* 2004.

[11] Soong, F. K., Lo, W. K. and Nakamura, S., "Optimal Acoustic and Language Model Weights for Minimizing Word Verification Errors," *Proc. ICSLP,* 2004.

[12] Lo, W. K., Soong, F. K. and Nakamura, S., "Robust Verification of Recognized Words in Noise," *Proc. ICSLP,* 2004.

[13] Ortmanns, S., Ney, H. and Aubert, X., "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition", *Computer Speech and Language*, Vol.11, pp.43-72, 1997.

[14] Qian, Y., Lee Tan and Li, Y. J., "Overlapped Di-tone Modeling for Tone Recognition in Continuous Cantonese Speech", *Proc. Eurospeech,* pp.1845-1848, 2003.

[15] CUCorpora: Cantonese Spoken Language Resources, 2001. http://dsp.ee.cuhk.edu.hk/speech/.

[16] Choi, W.N., Wong, Y.W., Lee Tan and Ching, P.C., "Lexical Tree Decoding with a Class-based Language Model for Chinese Speech Recognition", *Proc. ICSLP,* pp.174-177, 2000.

[17] Nelder, J. A. and Mead, R., "A Simplex Method for Function Minimization", *Computer Journal*, 7:308-313, 1965.

[18] Qian, Y. "Use of Tone Information in Cantonese LVCSR based on Generalized Character Posterior Probability Decoding", Ph. D Dissertation, The Chinese University of Hong Kong, 2005.