Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 22 (2008) 360–373
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Tone-enhanced generalized character posterior probability (GCPP) for Cantonese LVCSR

Yao Qian [a,b,*], Frank K. Soong [a,b], Tan Lee [b]

[a] Microsoft Research Asia, 5th Floor Beijing Sigma Center, No.49, Zhichun Road, Haidian District, Beijing 100080, PR China
[b] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, PR China

Received 29 November 2006; received in revised form 16 October 2007; accepted 17 December 2007
Available online 28 December 2007

Abstract

Tone-enhanced generalized character posterior probability (GCPP), a generalized form of posterior probability at sub-word (Chinese character) level, is proposed as a rescoring metric for improving Cantonese LVCSR performance. GCPP is computed by tone score along with the corresponding acoustic and language model scores. The tone score is output from a supra-tone model, which characterizes not only the tone contour of a single syllable but also that of adjacent ones and significantly outperforms other conventional tone models. 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 tone-enhanced GCPP, 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.

Keywords: Tone modelling; Posterior probability; Decoding criterion; Cantonese LVCSR

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. Thus, in such cases it would be more appropriate to use a cost function $\lambda(w_1^M, \hat{w}_1^N)$ to weight each sentence error. Here, we use $w_1^M = w_1, w_2, \ldots, w_M$ and $\hat{w}_1^N = \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_N$ to represent the hypothesized word string and candidate word string, respectively. The expected cost or risk associated with selecting w_1^M is defined as

* Corresponding author. Address: Microsoft Research Asia, 5flr Beijing Sigma Center, No.49, Zhichun Road, Haidian District, Beijing 100080, PR China. Tel.: +86 10 58965796; fax: +86 10 88099726.
E-mail addresses: yaoqian@microsoft.com (Y. Qian), frankkps@microsoft.com (F.K. Soong), tanlee@ee.cuhk.edu.hk (T. Lee).

0885-2308/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2007.12.003

$$R(w_1^M|x_1^T) = E[\lambda(w_1^M, \hat{w}_1^N)] = \sum_{\hat{w}_1^N} \lambda(w_1^M, \hat{w}_1^N) p(\hat{w}_1^N|x_1^T) \tag{1}$$

where $p(\hat{w}_1^N|x_1^T)$ is the posterior probability of word string $\hat{w}_1^N$ given the acoustic observations $x_1^T = x_1, x_2, \ldots, x_T$. The decision of speech recognition can be based on the minimization of the expected cost, i.e.

$$w_1^{*M} = \arg\min_{M, w_1^M} \sum_{\hat{w}_1^N} \lambda(w_1^M, \hat{w}_1^N) p(\hat{w}_1^N|x_1^T) \tag{2}$$

The MAP is in fact a special case of the minimum expected cost decision where a cost function 0 is assigned for two completely matched word strings and 1, otherwise. Eq. (2) can be rewritten as

$$w_1^{*M} = \arg\min_{M, w_1^M} \sum_{\substack{\hat{w}_1^N \\ \hat{w}_1^N \neq w_1^M}} p(\hat{w}_1^N|x_1^T) = \arg\max_{M, w_1^M} p(w_1^M|x_1^T) \tag{3}$$

MAP based decoding has been widely adopted in speech recognition because sentence with maximum posterior probability can be efficiently found by using Bayes rule and Viterbi search.

Many studies have been done on how to train a recognizer or perform search in recognition to optimize such a measure. For example, the cost function $\lambda(w_1^M, \hat{w}_1^N)$ of the Levenshtein (string edit) distance between two word strings w_1^M and $\hat{w}_1^N$, can be used to minimize the expected word error rate and it was proposed as the optimal search criterion for speech recognition (Stolcke et al., 1997; Mangu et al., 2000; Evermann and Woodland, 2000; Goel and Byrne, 2000). Estimation of word posterior probability and determination of the sentence with minimum expected word error were investigated for N-best output (Stolcke et al., 1997). They were also applied to a word graph (Mangu et al., 2000), where multiple string alignment instead of pairwise string alignment was adopted. In Goel and Byrne (2000), 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 (Wessel et al., 2000; Fetter et al., 1996; Neti et al., 1997).

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 (Weintraub, 1995; Wessel et al., 2001). Generalized posterior probability (Soong et al., 2004a) 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 (Soong et al., 2004b; Lo et al., 2004).

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 which is shared among many Chinese dialects, including the official spoken language, Mandarin or "Putonghua" in China. Each character is pronounced as a tonal 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 final performance of Chinese LVCSR is usually measured by character error rate (CER), rather than the word error rate.

There have been numerous studies on automatic tone recognition for Chinese ASR. Approaches to the subject fall into two major categories, namely, embedded tone modeling and explicit tone recognition. In embedded tone modeling, tone-related features such as F0 (the fundamental frequency) are included as extra components in the short-time feature vectors and consequently the acoustic models become tone-dependent (Chen et al., 1997; Huang and Seide, 2000; Wong and Chang, 2001; Wang et al., 2006). In this way, tone