Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 223–235
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Joint acoustic and language modeling for speech recognition

Jen-Tzung Chien *, Chuang-Hua Chueh

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan, ROC

Received 20 May 2008; received in revised form 21 September 2009; accepted 19 October 2009

Abstract

In a traditional model of speech recognition, acoustic and linguistic information sources are assumed independent of each other. Parameters of hidden Markov model (HMM) and *n*-gram are separately estimated for maximum *a posteriori* classification. However, the speech features and lexical words are inherently correlated in natural language. Lacking combination of these models leads to some inefficiencies. This paper reports on the joint acoustic and linguistic modeling for speech recognition by using the acoustic evidence in estimation of the linguistic model parameters, and vice versa, according to the maximum entropy (ME) principle. The discriminative ME (DME) models are exploited by using features from competing sentences. Moreover, a mutual ME (MME) model is built for *sentence posterior probability*, which is maximized to estimate the model parameters by characterizing the dependence between acoustic and linguistic features. The *N*-best Viterbi approximation is presented in implementing DME and MME models. Additionally, the new models are incorporated with the high-order feature statistics and word regularities. In the experiments, the proposed methods increase the sentence posterior probability or model separation. Recognition errors are significantly reduced in comparison with separate HMM and *n*-gram model estimations from 32.2% to 27.4% using the MATBN corpus and from 5.4% to 4.8% using the WSJ corpus (5K condition).

Keywords: Hidden Markov model; *n*-Gram; Conditional random field; Maximum entropy; Discriminative training; Speech recognition

1. Introduction

Speech recognition focuses on searching for the most likely word sequence $\widehat{W}$ from test speech X. The maximum *a posteriori* (MAP) decoding is performed by finding

$$\widehat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p_{\Lambda}(X|W)p_{\Gamma}(W) \tag{1}$$

where $p_{\Lambda}(X|W)$ denotes the acoustic likelihood given the hidden Markov model (HMM) Λ and $p_{\Gamma}(W)$ is the prior word probability given the *n*-gram model Γ. Statistical models Λ and Γ play an important role in speech recognition. Maximum likelihood (ML) is a popular criterion for parameter estimation. However, higher likelihood does not guarantee better classification. Minimum classification error (MCE) and maximum mutual information (MMI) criteria were presented for discriminative training of HMMs (Bahl et al., 1986; Juang and Katagiri, 1992; Normandin et al., 1994) and *n*-grams (Kuo et al., 2002).

The maximum entropy (ME) method (Jaynes, 1957; Della Pietra et al., 1997) is attractive for establishing model distribution with maximum randomness subject to certain constrains. ME estimation was exploited for language modeling (Berger et al., 1996; Rosenfeld, 1996), and extended for direct acoustic modeling (Kuo and Gao, 2004). The merit of ME modeling is the capabilities of merging non-independent, asynchronous and overlapping features into a joint probability model. The information sources of trigger pairs, association patterns and semantic topics were incorporated in language models (Rosenfeld, 1996; Khudanpur and Wu, 2000; Chien, 2006). Kuo and Gao (2004) combined the asynchronous sequences of speech observations and HMM states into acoustic model, which corresponded to the maximum entropy Markov

* Corresponding author. Tel.: +886 6 2757575x62532; fax: +886 6 2747076.
E-mail address: jtchien@mail.ncku.edu.tw (J.-T. Chien).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.10.003

model (MEMM) (McCallum et al., 2000). The conditional random field (CRF) (Lafferty et al., 2001) has been proposed to perform global normalization rather then the local normalization in an MEMM. The model parameters were estimated by maximizing the conditional likelihood. Liu et al. (2006) investigated the properties of HMM, MEMM and CRF for sentence boundary detection. Sha and Pereira (2003) applied CRF for shallow parsing where different training methods were evaluated by the metric of labeling accuracy. CRF was also explored for phone recognition by adopting various phonetic attributes rather than traditional ceptral features as the input features (Morris and Fosler-Lussier, 2006). Due to the discriminatively trained attributes, the phone recognition accuracy was improved.

Previous studies assumed acoustic and linguistic models (Λ, Γ) were independent, and estimated them individually by different criteria. A scaling factor was required to balance the acoustic and linguistic scores, and seen as an important parameter in a large vocabulary continuous speech recognition (LVCSR) system although some other decoding parameters e.g. beam width, insertion penalty, etc., are also tunable for LVCSR. Two weaknesses are incurred. First, considering the hierarchical matching from phonetic level to lexical level, the assumption of model independence is unrealistic for obtaining the global optimum of concerned criteria. Second, the scaling factor is sensitive to the changing environments and domains. A development set should be prepared to tune the scaling factor. In (Beyerlein, 1998), a discriminative combination of acoustic model and language model was presented by adapting the scaling factor rather than the models themselves according to different discriminative criteria.

Recently, Gunawardana et al. (2005) presented the hidden CRF (HCRF) for phone classification, where the HMM parameters and phone unigrams were combined in a log linear model for joint optimization. Model dependency between acoustic and linguistic features was first investigated. Nonetheless, HCRF was not feasible for LVCSR because the marginalization over state sequences was intractable beyond phone classification. Although the ME and CRF algorithms were derived from different objective functions, they complied with the same log linear model in parameter estimation. Additionally, Quattoni et al. (2007) developed a discriminative hidden-state CRF model for visual recognition tasks. This model was equivalent to the HCRF model and outperformed the HMM and CRF models for gesture recognition. The advantage of joint discriminative learning of latent variables and observations was demonstrated.

In this study, we recast HCRF from the viewpoint of ME principle, and incorporate the higher-order acoustic and linguistic features in calculation of sentence posterior probability for LVCSR. The evolution of ME models is presented. First, the acoustic features from competing word candidates are adopted to estimate the discriminative ME (DME) language model (Chueh et al., 2005). Acoustic discrimination information is embedded in linguistic parameters. The relationship of DME model to the other models is illustrated. Moreover, a mutual ME (MME) model is built by combining different sources of acoustic and linguistic features (Chueh and Chien, 2006). This modularized framework is initialized from the individual HMM and n-gram models, and generalized to characterize high-order acoustic and language regularities. In particular, HMM and n-gram parameters are merged for joint optimization. The LM scaling factor is not required in decoding algorithm. The discriminative training is performed by using the N-best paradigm for word recognition. This work investigates how different features affect objective function and speech recognition accuracy by performing experiments on broadcast news data.

2. Maximum entropy modeling

An ME method focuses on completely modeling what is known, and carefully avoiding assuming anything that is not known (Jaynes, 1957; Berger et al., 1996). All information sources serve as constraints to be imposed to infer the model with the highest entropy. This approach is more elaborate than information combination by linear interpolation, since the ME model optimally combines the information sources in a consistent fashion. The ME acoustic model has been developed by directly modeling the HMM states and observation Gaussian identities (Kuo and Gao, 2004). HCRF model has been examined to characterize phone labels, the HMM states and Gaussian statistics (Gunawardana et al., 2005; Lafferty et al., 2001). This study explores the joint acoustic and linguistic modeling based on ME principle and investigates the equivalence to the HCRF model.

2.1. ME principle

Given a probability model $p(y)$ for random variable y, ME method restricts the model to be consistent with all information sources, and simultaneously makes the model distribution as uniform as possible. Let $f_1(y), \ldots, f_F(y)$ denote the feature functions defined by

$$f_i(y) = \begin{cases} 1, & \text{if } y \text{ matches feature } i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Since the true model $p(y)$ encapsulates these features, the expected feature functions using true distribution $p(y)$ and empirical distribution $\tilde{p}(y)$ should satisfy the equality

$$E_p[f_i] = E_{\tilde{p}}[f_i], \qquad \text{for } i = 1, \ldots, F \tag{3}$$

where the true expectation is given by $E_p[f_i] = \sum_y p(y) f_i(y)$, and the empirical expectation is calculated from training samples $\{y_t,\ t = 1, \ldots, T\}$ by $E_{\tilde{p}}[f_i] = (1/T)\sum_{t=1}^{T} f_i(y_t)$. The ME model is calculated by maximizing the entropy generated by $p(y)$ such that the constraints in (3) are satisfied. The extended objective function is formed by