Options for Modelling Temporal Statistical Dependencies in an Acoustic Model for ASR

Volker Leutnant, Reinhold Haeb-Umbach
Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
Email: {leutnant,haeb}@nt.uni-paderborn.de

Abstract

In this paper we consider the combination of hidden Markov models based on Gaussian mixture densities (GMM-HMM) and linear dynamic models (LDM) as the acoustic model for automatic speech recognition systems. In doing so, the individual strengths of both models, i.e. the modelling of long-term temporal dependencies by the GMM-HMM and the direct modelling of statistical dependencies between consecutive feature vectors by the LDM, are exploited. Phone classification experiments conducted on the TIMIT database indicate the prospective use of this approach in continuous speech recognition.

Index Terms: acoustic model, linear dynamic model, statistical model combination, phone classification

Introduction

Traditionally, automatic speech recognition systems are based on hidden Markov models with Gaussian mixtures modelling the state-conditioned feature vector distribution. The inherent assumption of conditional independence, stating that a feature vector's likelihood solely depends on the current HMM state, makes the search computationally tractable, nevertheless has also been identified to be a major reason for the lack of robustness. Linear dynamic models have been proposed to overcome this weakness by employing a hidden dynamic state process underlying the observed feature vectors. Though performance of LDMs on phone classification tasks has been shown to be superior to that of an equivalent static model (i.e. single-state monophone HMMs with unimodal full covariance Gaussian emission density), this approach still cannot compete with the established acoustic models (i.e. multi-state triphone HMMs with multimodal diagonal covariance Gaussian emission densities) when it comes to continuous speech recognition [1].

Nevertheless, it is believed that LDM and GMM-HMM have complementary strengths to be exploited to achieve an overall gain in performance. Thus, on the way towards a hybrid decoder architecture, combination of information provided by the LDM and GMM-HMM, respectively, will be examined in this paper.

Acoustic Models

Speech recognition usually considers the observed feature vectors to be generated by a stochastic process. Since this process is unknown, acoustic modelling aims at finding a representation of it that is a) closely reflecting the (seen and unseen) data, yet b) being computational tractable. The acoustic models considered in this paper are the Gaussian mixture model based HMM and the LDM.

GMM-HMM

The hidden Markov model is the most popular approach to model the observed features. By introducing a hidden, discrete-valued state process underlying the observation process, the likelihood of a sequence of observations $\mathbf{y}_1^\tau$ given a hypothesised phone ω_k is

$$p(\mathbf{y}_1^\tau|\omega_k) \approx \max_{q_1^\tau} \prod_{t=1}^{\tau} p(\mathbf{y}_t|q_t,\omega_k) P(q_t|q_{t-1},\omega_k),$$

where the maximization in the above VITERBI-approximation has to be carried out over all possible sequences q_1^τ of hidden states. The state-conditioned feature vector distribution $p(\mathbf{y}_t|q_t,\omega_k)$ is usually modelled as a mixture of M (diagonal covariance) Gaussians

$$p(\mathbf{y}_t|q_t=j,\omega_k) = \sum_{i=1}^{M} c_{i,j,k}\, \mathcal{N}\left(\mathbf{y}_t; \boldsymbol{\mu}_{i,j,k}, \boldsymbol{\Sigma}_{i,j,k}\right),$$

with weights $c_{i,j,k}$, means $\boldsymbol{\mu}_{i,j,k}$ and covariances $\boldsymbol{\Sigma}_{i,j,k}$.

LDM

Linear dynamic models have been proposed as an alternative acoustic model for phone classification and recognition [1]. The LDM system is based on a hidden, linear, autoregressive, continuous-valued state process underlying the observation process. A linear measurement equation relates the hidden state process to the observation. The likelihood of a sequence of observations $\mathbf{y}_1^\tau$ given a hypothesised phone ω_k is

$$p(\mathbf{y}_1^\tau,\omega_k) = \prod_{t=1}^{\tau} p(\mathbf{y}_t|\mathbf{y}_1^{t-1},\omega_k)$$

where $p(\mathbf{y}_t|\mathbf{y}_1^{t-1},\omega_k)$ can be solved analytically if state process and measurement equation are linear and driven by (uncorrelated) Gaussian noises, resulting in the standard KALMAN filtering.

Phone Classification

Given a phone alignment of a sequence of N phones Ω_1^N ($\Omega_n \in \{\omega_1,...,\omega_K\}$) and a corresponding observation vector sequence $\mathbf{y}_1^T$ of length T, phone classification looks for the most probable phone sequence given the feature vector sequence and the alignment. The alignment can be expressed by a sequence of phone start and end times $\{t_0^{N-1}+1, t_1^N\}$, with t_0=0 and t_N=T, or equivalently by a sequence of phone durations l_1^N.

Neglecting "cross-phone" dependencies on the feature vector level, the posterior probability of a phone sequence can be decomposed into a (scaled) m-gram language model prior, a (scaled) duration model likelihood and the acoustic likelihood.

Bayes' Decision Rule is thus given by

$$\widehat{\Omega}_1^N = \arg\max_{\tilde{\Omega}_1^N} P(\tilde{\Omega}_1^N|\mathbf{y}_1^T, t_0^{N-1}, l_1^N)$$
$$= \arg\max_{\tilde{\Omega}_1^N} \prod_{n=1}^{N} \underbrace{p(\mathbf{y}_{t_{n-1}+1}^{t_n}|\tilde{\Omega}_n)}_{\substack{\text{acoustic}\\ \text{model}}} \underbrace{p(l_n|\tilde{\Omega}_n)^{\gamma}}_{\substack{\text{duration}\\ \text{model}}} \underbrace{P(\tilde{\Omega}_n|\tilde{\Omega}_{n-m+1}^{n-1})^{\beta}}_{\substack{\text{language}\\ \text{model}}},$$

which can be solved by the VITERBI-Algorithm.

Statistical Model Combination

Finding the optimal phone sequence requires information from the acoustic, the duration and the language model to be gathered. Traditionally, the acoustic model is composed of GMM-HMMs. While the GMM-HMMs are capable of modelling long-term temporal dependencies, LDMs allow for direct temporal statistical dependencies between consecutive feature vectors to be modelled. Utilising LDMs as an acoustic model for phone classification has intensively been studied by Frankel [1], who already noted that both models have their strengths and weaknesses and considered their combination by means of weighted averaging their individual likelihoods for a given segment and hypothesised phone. In general, statistical combination of multiple acoustic models can happen on either the "likelihood level" or the "phone posterior level". This paper focuses on the latter combination approach, with phone posterior probabilities being computed on wordgraphs [2]. The probability of a sequence of phones can thus be approximated by the product of posterior probabilities of involved phones. The following combination methods have been examined:

- GMM-HMM/LDM: always select either the GMM-HMM or the LDM; this experiment gives the baselines for all selection and combination methods following;
- Minimum Entropy: select the acoustic model with minimum entropy on the current segment [3];
- Max/Min: select that model to support a phone hypotheses giving the highest/lowest posterior probability for it;
- Sum/Product: the support for a phone is the weighted sum/exponentially weighted product of the individual posterior probabilities;
- Inverse Entropy: sum rule with weights inversely proportional to the entropy of the acoustic models [3];
- Entropy-based DS: DEMPSTER-SHAFER model combination [3]; weights of the ignorance models are based on the entropy of the acoustic models;

Experimental Results

Based on phonetic annotations given by the TIMIT corpus [4], training of 61 context-independent LDM and GMM-HMM phone models has been carried out under the expectation maximization framework, with the LDMs and GMM-HMMs based on a linear, autoregressive state process of order 1 and a 3–state HMM with linear topology, respectively. A log-Gaussian duration model and an unsmoothed phone bigram language model have been build on the same data. With standard 39-dimensional MFCC+Δ+Δ^2 feature vectors, likelihoods for each segment of an utterance have been computed and stored in wordgraphs, followed by the computation of phone posterior probabilities for the LDM and the GMM-HMM. Optimal training and test parameters have been determined on the TIMIT development set for a GMM with 20 mixtures, giving an LDM state dimension of 12 and duration and language model scaling factors of 8.

The introduced combination approaches have been evaluated on the TIMIT test set for varying numbers of mixtures in the GMM-HMM. Classification results on a collapsed phone set of cardinality 39 are displayed below.

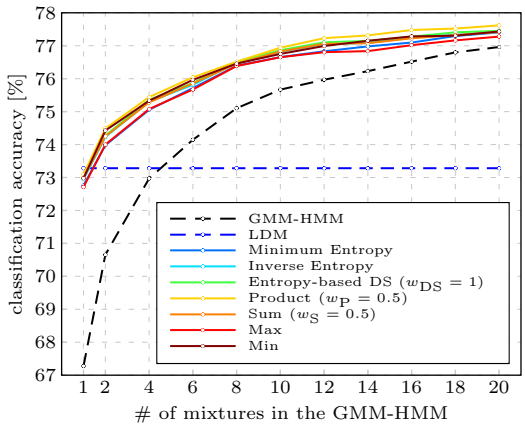

Figure 1: Classification results for a varying number of mixtures in the GMM-HMM

With the number of mixtures in the GMM-HMM greater or equal to 2, all combination methods yield significant improvements over the best individual model's classification accuracy, with the simple product rule giving the overall best results. The inverse entropy rule and the DEMPSTER-SHAFER rule perform equally well. Choosing the weights in the sum, product and DEMPSTER-SHAFER rule different from the default weights used to create figure 1 (w_S=0.5, w_P=0.5, w_{DS}=1) further improves the classification accuracies. However, optimization has to be carried for each GMM-HMM (i.e. with respect to the number of states and mixtures) individually.

Conclusions

In this paper we have examined the combination of GMM-HMM and LDM acoustic models for phone classification by means of phone posterior probability combination. Computation of phone posterior probabilities has been carried out on wordgraphs. Significant improvements obtained by all introduced combination methods motivate further exploration of the acoustic model combination and its application to continuous speech recognition.

Acknowledgement

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/6-1.

References

[1] J. Frankel, "Linear dynamic models for automatic speech recognition," Ph.D. dissertation, University of Edinburgh, Edinburgh, UK, 2003.

[2] F. Wessel, "Word posterior probabilities for large vocabulary continuous speech recognition," Ph.D. dissertation, RWTH Aachen, Aachen, Germany, 2002.

[3] F. Valente, "Multi-stream speech recognition based on dempster-shafer combination rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, 2010.

[4] W. M. Fisher, G. R. Doddington, Goudie-Marshall, and K. M., "The darpa speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.