

Cross-stream Observation Dependencies for Multi-stream Speech Recognition

Özgür Çetin, Mari Ostendorf

Signal, Speech, and Language Interpretation Laboratory
Department of Electrical Engineering, University of Washington, Seattle
{cozgur, mo}@ee.washington.edu

Abstract

This paper extends prior work in multi-stream modeling by introducing cross-stream observation dependencies and a new discriminative criterion for selecting such dependencies. Experimental results combining short-term PLP features with long-term TRAP features show gains associated with a multi-stream model with partial state asynchrony over a baseline HMM. Frame-based analyses show significant discriminant information in the added cross-stream dependencies, but so far there are only small gains in recognition accuracy.

1. Introduction

Multi-stream models of speech combine information extracted from multiple sources to improve recognition. The information sources are typically feature sets which have been derived either from audio signal itself, as in sub-band processing [1], or from some associated signal, such as a video of the speaker's mouth [2]. The multi-stream model provides a general framework for signal processing techniques that decompose a time-series into various scale- and/or frequency-localized components, and for multi-modal recognition applications. The multi-stream approach has gained increasing popularity in recent years, because it can deal with multiple sources of information without making over-restrictive assumptions about how the separate feature streams relate to the underlying word sequence and to each other. Also, the performance of multi-stream models degrades gracefully against interference affecting a subset of streams, e.g. narrowband noise in a multi-band system.

Previous work in multi-stream models has mainly focused on controlling asynchrony among streams and on methods of combining probability scores of streams. The observation streams depend on each other through the coupling of hidden state sequences, assuming conditional independence of observations given the respective state sequences [3]. This assumption discards potentially discriminative direct dependencies among feature streams. Here, we extend the multi-stream model to allow for such direct dependencies between the observation streams. Even though an arbitrary selection of dependencies always increases the descriptive power of the model, i.e. training likelihood, this does not necessarily increase its accuracy. Only dependencies which make the model

more discriminative should be chosen. In this work, we investigate whether such discriminative information exists, through calculating information-theoretic quantities between feature streams, and then choose sparse dependencies according to those quantities. Experiments are in the context of a cross-domain speech recognition task in which training data consists of unrestricted vocabulary conversational speech utterances whereas testing data is number sequences. We use conventional short-term spectral features, perceptual linear predictive (PLP) coefficients, and long-term tandem-based TRAPs (Temporal Patterns) as observation streams. Unlike previous work [4] that finds no evidence of loss of discriminative information by the state conditional independence assumption of chains in multi-band speech recognition, we show that there exists such information between PLPs and TRAPs that is not captured by state coupling alone.

The organization of this paper is as follows. The multi-stream model and our extension of it to allow cross-stream dependencies are described in Sections 2 and 3. In Section 4, we describe the automatic dependency selection criteria for choosing cross-stream dependencies. The task description and experimental results are given in Section 5, and we summarize findings in Section 6.

2. Multi-stream Model

A hidden Markov model (HMM) characterizes the joint distribution of a length T time-series $\{o_t\}_{t=0}^{T-1}$ through an underlying hidden state sequence $\{s_t\}_{t=0}^{T-1}$,

$$p(\{o_t\}_{t=0}^{T-1}, \{s_t\}_{t=0}^{T-1}) \equiv \prod_{t=0}^{T-1} p(s_t | s_{t-1}) p(o_t | s_t)$$

in which one assumes that the state sequence is first-order Markov, s_{-1} is a start state, and observations are independent of everything else given their respective states. Suppose that instead of a single observation time-series, one has two time-series, $\{o_t^1\}_{t=0}^{T-1}$ and $\{o_t^2\}_{t=0}^{T-1}$. One approach concatenates the two feature sequences together to form a single higher-dimension sequence modeled by an HMM, forcing the underlying state sequence to be the same for both streams. An alternative approach is to assume independence of the two sequences and model each

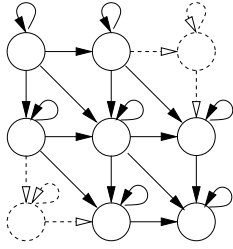


Figure 1: *Joint state topology of two coupled HMMs with partial asynchrony; disallowed states are in dashes.*

with separate HMMs. The first approach fails to capture asynchrony in the timing of features corresponding to the events of the original process, while the second assumes that there is no common event process and ignores any information in the joint dynamics of the two streams. A compromise between the two extremes is obtained by coupling the two HMMs through their state sequences:

$$p(\{o_t^1\}_{t=0}^{T-1}, \{s_t^1\}_{t=0}^{T-1}, \{o_t^2\}_{t=0}^{T-1}, \{s_t^2\}_{t=0}^{T-1}) \equiv \prod_{t=0}^{T-1} p(s_t^1 | s_{t-1}^1) p(o_t^1 | s_t^1) p(s_t^2 | s_{t-1}^2, s_t^1) p(o_t^2 | s_t^2) \quad (1)$$

where $\{s_t^{1,2}\}_{t=0}^{T-1}$ are the state sequences underlying each observation sequence. In this model, the observation sequences are not marginally independent, but they are *conditionally* independent given their respective state sequences. Various degrees of synchrony between the two sequences are obtained by restricting allowable transitions in the product state space of the two streams, as shown in Figure 1. The computational cost of inference is $O(|\mathcal{S}|^3 T)$, which compares to $O(|\mathcal{S}|^2 T)$ for HMMs.

The state conditional observation distributions are usually modeled by a mixture of Gaussians,

$$p(o_t | s_t = i) = \sum_l \alpha_{il} \mathcal{N}(o_t; \mu_{il}, \Sigma_{il})$$

where we have suppressed stream index superscripts to simplify notation, and $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian random vector with mean μ and covariance matrix Σ .

3. Cross-stream Observation Dependencies

In the coupled HMMs described above, any dependence between the two observation streams is mediated through the hidden state variables. This assumption might be too simplistic for speech modeling given that $o_t^{1,2}$ are derived from the same speech signal even if using different feature extraction methods. In past work, data-driven corpus studies have consistently shown that similar temporal conditional independence assumptions of HMMs do not hold, and there lies significant information between current and surrounding observations conditional on the underlying state sequence label [5]. Here we question the

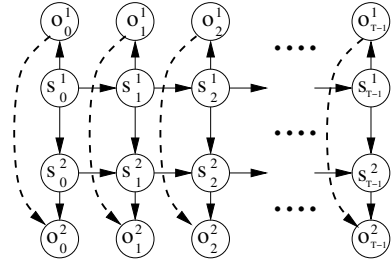


Figure 2: *A graphical model depiction of the multi-stream model; cross-stream dependencies have dashed lines.*

assumption of conditional independence of observations extracted from time slices centered around the same time using different windows and signal processing methods. Experiments in Section 5 will show that indeed this assumption does not always hold, and that there is discriminative information in the dependence between observation streams.

We extend the state-coupled HMM multi-stream model to allow direct dependencies from one observation sequence to another, say from sequence 1 to sequence 2, which amounts to replacing $p(o_t^2 | s_t^2)$ by $p(o_t^2 | o_t^1, s_t^2)$ in the factorization of the joint density of observations and states in Equation 1. A graphical model illustration of the extended dependencies is in Figure 2. Due to their mathematical tractability, we use mixtures of linear conditional Gaussians to model the dependence of o_t^2 on o_t^1 ,

$$p(o_t^2 | o_t^1, s_t^2 = i) = \sum_l \alpha_{il} \mathcal{N}(o_t^2; A_{il} o_t^1 + \mu_{il}, \Sigma_{il})$$

where A_{il} is a $d_2 \times d_1$ regression matrix of o_t^2 on o_t^1 , and d_k is the dimension of o_t^k . Maximum likelihood (ML) estimation is straightforward using the Expectation-Maximization (EM) algorithm for updating the regression matrices and other model parameters. The additional cross-stream observation dependencies do not significantly increase the computational cost of inference (decoding) since $\{o_t^{1,2}\}$ are always observed. However, arbitrary dependencies increase the number of parameters to be estimated from finite amounts of training data, as well as storage costs of various sufficient statistics that are used in EM updates which is similar to the costs associated with the training of full-covariance Gaussians. Hence, only a judicious choice of dependencies can lead to reliable parameter estimates and more accurate classifiers. The next section describes a method of choosing such discriminative sparse dependencies.

4. Dependency Selection Criteria

Suppose that one has a class variable S , and two features, X and Y , which are associated with S and from which S is to be predicted. The true generative model, $p(X, Y, S)$, is unknown. We assume a model, $q(X, Y, S)$, to be esti-

mated from data using ML training. Given that S is the class attribute to be predicted, we assume that $S \rightarrow X$ and $S \rightarrow Y$ dependencies are always selected. When does the modeling of a $X \rightarrow Y$ dependency increase the accuracy of classifications made according to $q(S|X, Y)$? The different assumptions factorize the joint distribution of $\{X, Y, S\}$ according to,

$$\begin{aligned} q_a(x, y, s) &= q(s)q(x|s)q(y|s), \\ q_b(x, y, s) &= q(s)q(x|s)q(y|x, s). \end{aligned}$$

Notice that $q_a(\cdot)$ is a submodel of $q_b(\cdot)$, and hence the ML criteria will always select the $X \rightarrow Y$ dependency, and is useless for assessing its discriminative power.

To assess the discriminative power of the $X \rightarrow Y$ dependency, ideally one wants to investigate how $q_a(\cdot)$ and $q_b(\cdot)$ compare in terms of the error rate of their maximum a posteriori probability decisions,

$$\pi_{a,b} \equiv 1 - E_p[p(\alpha_{a,b}(X, Y)|X, Y)]$$

where $\alpha_{a,b}(X, Y) \equiv \operatorname{argmax}_s q_{a,b}(S|X, Y)$, and $E_p[\cdot]$ denotes expectation with respect to $p(\cdot)$. The error rate above does not lend itself to further manipulation due to its nonsmooth form. Instead, we will investigate the cross-entropy between true and assumed distributions:

$$H_p(q_{a,b}) \equiv -E_p[\log q_{a,b}(S|X, Y)]$$

which is essentially the negative conditional log-probability of class labels or, equivalently, the negative of the objective function for maximum mutual information estimation. This quantity can be thought of as a measure of classifier accuracy or “confidence”: smaller $H_p(q)$ means a better classifier. A straightforward manipulation of $H_p(q_a)$ and $H_p(q_b)$, under the assumption that their components are estimated with the ML criteria, yields $H_p(q_b) = H_p(q_a) - G(X; Y|S)$ where

$$\begin{aligned} G(X; Y|S) &\equiv I_p^b(X; Y|S) - I_p^b(X; Y) + I_p^a(X; Y) \\ I_p^b(X; Y|S) &\equiv E_p[\log(q_b(X|Y, S)/q_b(X|S))] \\ I_p^b(X; Y) &\equiv E_p[\log(q_b(X|Y)/q_b(X))] \end{aligned}$$

and $I_p^a(X; Y)$ is defined similarly. Hence, the $X \rightarrow Y$ dependency decreases the entropy of the posterior of the class variable only if $G(X; Y|S)$ is positive. Roughly speaking, this corresponds to having the conditional mutual information (MI) between X and Y be more than the unconditional MI, i.e. X and Y are more informative about each other when S is known. The implication of this result for structure learning is that the $X \rightarrow Y$ dependency should be chosen if $G(X; Y|S)$ is large and positive. A large negative value means the dependency is redundant and likely to *decrease* classification accuracy. The first two terms in $G(X; Y|S)$ together correspond to the explaining away residual measure which was first derived in [5] under slightly different assumptions.

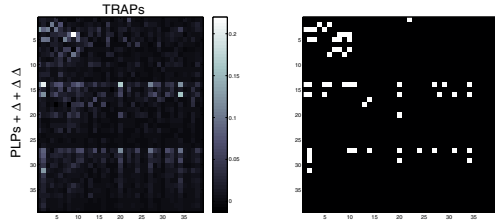


Figure 3: G measure between PLPs and TRAPs estimated from frames which have been aligned to the final states of triphones of phone “aa”. The bitmap of coordinates where measure is $\geq .05$ is shown on the right.

The problem of selecting discriminative cross-stream observation dependencies is exactly analogous to this problem, letting S correspond to a state in the product state space, and o_t^1 and o_t^2 correspond to X and Y , respectively. We use the above result by estimating $G(X; Y|S)$ for each pair of features in o_t^1 and o_t^2 , given a large speech corpus, and choose candidate links with measures above a threshold. (See Section 5 for details.) As an example, we have depicted G measure between PLPs and TRAPs conditional on the third state of phone “aa” in Figure 3, revealing candidate dependencies and showing that there exists significant discriminative information between the two feature streams which cannot be captured through state coupling. Note that most pairwise measures are around zero – evidence that arbitrary selection of dependencies might not improve the performance.

5. Experiments

Task. We have applied multi-stream modeling to a cross-domain speech recognition task to assess generalizability of the models. The training data set includes roughly 16 hours of speech, including conversational speech drawn from the Switchboard and Callhome corpora (9.6 hours), and read speech from the Macrophone corpus (6.4 hours). The testing data is roughly 1.2 hours of spoken number sequences from the Numbers95 corpus. Even though the decoding vocabulary (35 words) and its phone coverage¹ is limited, reliable estimation of phones from open vocabulary training data requires training of context-dependent models for all phones. Hence, the training paradigm is similar to that for HMMs in large vocabulary speech recognition and requires use of state-tying.

Features. Two feature streams are used: PLP and TRAP coefficients. We compute 12 PLP coefficients and log-energy from 25ms windows of speech with a 100Hz frame rate, plus their first and second differences (39 dimensions total). Per-side mean subtraction and variance normalization are used. The TRAP features are extracted as follows: 15 critical band energies derived from 1s win-

¹Only 28 of 48 training phones appear in the decoding dictionary.

dows of speech are used to train 15 separate neural networks on 47 phone targets, and the outputs of the individual networks are merged by another neural-network with 47 outputs omitting the final nonlinearity. For more details, see [6]. Here, we reduce the TRAP dimension to 39 through principal component analysis.

Multi-stream Model. We implemented a context-dependent phone-based multi-stream model with arbitrary state tying within each stream using the Graphical Models Toolkit (GMTK) [7]. Word-internal triphones are modeled through state coupled HMMs with and without cross-stream observation dependencies. The two streams are forced to synchronize at phone boundaries, but in between partial asynchrony of two streams is allowed using the state topology of Figure 1. The tying of states within each stream is obtained from decision tree clustering of triphones in separate HMMs for each feature stream. The total number of unique states for PLP and TRAP streams are 868 and 1105, respectively, and each state conditional observation density is modeled by a mixture of 16 diagonal Gaussians. The joint training of all parameters is done via the EM algorithm. Single-pass decoding is performed by finding the most likely word sequence in the joint state space of two streams.

Dependency Selection. The sparse dependencies of the PLP feature stream on TRAPs are selected through the G measure, which requires estimates of unconditional and composite state-specific MI values between two streams which are obtained as follows. First, for each pair of features in each combination of phone and state position (begin, middle, end), a mixture of Gaussians is fit to data that has been force-aligned to such combinations by an HMM system using concatenated PLP and TRAP features. The conditional and unconditional MIs are then obtained by invoking the law of large numbers on the respective distributions. Using the state-specific G measures, we selected at most two links per PLP coefficient for each phone and state position, if the measure is more than a threshold ($= .05$). Overall, 80% of all chosen links are received by the first 5 PLP coefficients and their first- and second-order differences, and boundary states received more links than middle states. The structure (but not the parameter) is shared by all states in the same state position in all triphones of a phone.

Results. We compare multi-stream models with and without cross-stream observation dependence to the baseline HMM system. For reference, we also include the word error rate (WER) of an HMM system using concatenated PLP and TRAP features. All experiments use a bigram language model trained on the Numbers95 training data, and the total number of parameters in each system are roughly the same. The results are depicted in Table 1. The standard multi-stream model improves over baseline

HMM and feature concatenation approaches when used with partial asynchrony. Adding cross-stream observation dependencies gives a small additional gain, though not statistically significant.

model	features	WER%
HMM	PLP	3.8
HMM (concat. features)	PLP+TRAPs	3.8
MULTI-STREAM	PLP+TRAPs	3.7
MULTI-STREAM+links	PLP+TRAPs	3.6

Table 1: WERs of various systems on Numbers95.

6. Summary and Future Work

In this work, we extended the state-coupled multi-stream model to allow for direct cross-stream observation dependencies, which we show to have discriminative information for speech recognition. In addition, we described a discriminative selection criterion to choose sparse dependencies. We implemented a context-dependent phone-based multi-stream model with PLP and TRAP feature streams, and showed that the standard multi-stream model with partial asynchrony improves over a baseline HMM. While the performance gain due to discriminatively selected links was small (not statistically significant), it may be that the cross-domain testing paradigm limits the gains from discriminatively trained techniques, a question that we will explore in future work.

Acknowledgments

The authors thank J. Bilmes of UW for helpful discussions and support with GMTK and MI estimation software, P. Jain and F. Grezl of OGI for their help with TRAPS, and B. Chen of ICSI for his help with the Numbers95 corpus.

7. References

- [1] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP*, pp. 426-429, 1996.
- [2] J. Luetin, G. Potamianos and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," *Proc. ICASSP*, pp. 169-172, 2001.
- [3] H. Nock and S. Young, "Loosely coupled HMMs for ASR," *Proc. ICSLP*, pp. 143-146, 2000.
- [4] N. Mirghafori and N. Morgan, "Transmissions and transitions: A study of two common assumptions in multi-band ASR," *Proc. ICASSP*, pp. 713-716, 2001.
- [5] J. Bilmes, "Buried Markov models for speech recognition," *Proc. ICASSP*, pp. 713-716, 1999.
- [6] H. Hermansky, S. Sharma and P. Jain, "Data-derived non-linear mapping for feature extraction in HMM," *Proc. ASRU Workshop*, 1999.
- [7] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," *Proc. ICASSP*, pp. 3916-3919, 2002.