Latent Mixture of Discriminative Experts for Multimodal Prediction Modeling

Derya Ozkan, Kenji Sagae and Louis-Philippe Morency
USC Institute for Creative Technologies
{ozkan,sagae,morency}@ict.usc.edu

Abstract

During face-to-face conversation, people naturally integrate speech, gestures and higher level language interpretations to predict the right time to start talking or to give backchannel feedback. In this paper we introduce a new model called Latent Mixture of Discriminative Experts which addresses some of the key issues with multimodal language processing: (1) temporal synchrony/asynchrony between modalities, (2) micro dynamics and (3) integration of different levels of interpretation. We present an empirical evaluation on listener nonverbal feedback prediction (e.g., head nod), based on observable behaviors of the speaker. We confirm the importance of combining four types of multimodal features: lexical, syntactic structure, eye gaze, and prosody. We show that our Latent Mixture of Discriminative Experts model outperforms previous approaches based on Conditional Random Fields (CRFs) and Latent-Dynamic CRFs.

1 Introduction

Face-to-face communication is highly interactive. Even when only one person speaks at a time, other participants exchange information continuously amongst themselves and with the speaker through gestures, gaze and prosody. These different channels contain complementary information essential to interpretation and understanding of human behaviors (Oviatt, 1999). Psycholinguistic studies also suggest that gesture and speech come from a single underlying mental process, and they are related both temporally and semantically (McNeill, 1992; Cassell and Stone, 1999; Kendon, 2004).

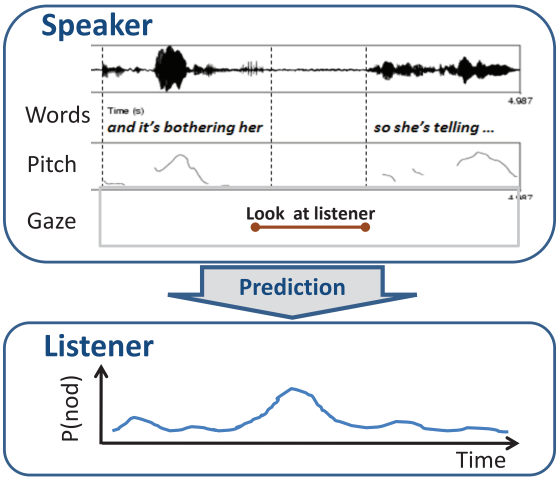


Figure 1: Example of multimodal prediction model: listener nonverbal backchannel prediction based on speaker's speech and eye gaze. As the speaker says the word *her*, which is the end of the clause (*her* is also the object of the verb *bothering*), and lowers the pitch while looking back at the listener and eventually pausing, the listener is then very likely to head nod (i.e., nonverbal backchannel).

A good example of such complementarity is how people naturally integrate speech, gestures and higher level language to predict when to give backchannel feedback. Building computational models of such a predictive process is challenging since it involves micro dynamics and temporal relationship between cues from different modalities (Quek, 2003). Figure 1 shows an example of backchannel prediction where a listener head nod is more likely. For example, a temporal sequence from the speaker where he/she reaches the end of segment (syntactic feature) with a low pitch and looks at the listener before pausing is a good opportunity for the listener to give nonverbal feedback (e.g., head nod). These prediction models have broad applicability, including the improvement of nonverbal behavior recognition, the synthesis of natural animations for robots and virtual humans, the training of cultural-specific nonverbal behaviors, and the diagnoses of social disorders (e.g., autism spectrum disorder).

In this paper we introduce a new model called Latent Mixture of Discriminative Experts (LMDE) which addresses some of the key issues with multimodal language processing: (1) temporal synchrony/asynchrony between modalities, (2) micro dynamics and (3) integration of different levels of interpretation. We present an empirical evaluation on nonverbal feedback prediction (e.g., head nod) confirming the importance of combining different types of multimodal features. We show that our LMDE model outperforms previous approaches based Conditional Random Fields (CRFs) and Latent-Dynamic CRFs.

2 Related Work

Earlier work in multimodal language processing focused on multimodal dialogue systems where the gestures and speech may be constrained (Johnston, 1998; Jurafsky et al., 1998). Most of the research in multimodal language processing over the past decade fits within two main trends that have emerged: (1) recognition of individual multimodal actions such as speech and gestures (e.g, (Eisenstein et al., 2008; Frampton et al., 2009; Gravano et al., 2007)), and (2) recognition/summarization of the social interaction between more than one participants (e.g., meeting analysis (Heylen and op den Akker, 2007; Moore, 2007; Murray and Carenini, 2009; Jovanovic et al., 2006)).

The work described in this paper can be seen from a third intermediate category where multimodal cues from one person is used to predict the social behavior of another participant. This type of predictive models has been mostly studied in the context of embodied conversational agents (Nakano et al., 2003; Nakano et al., 2007). In particular, backchannel feedback (the nods and paraverbals such as "uh-hu" and "mm-hmm" that listeners produce as someone is speaking) has received considerable interest due to its pervasiveness across languages and conversational contexts and this paper addresses the problem of how to predict and generate this important class of dyadic nonverbal behavior.

Several researchers have developed models to predict when backchannel should happen. In general, these results are difficult to compare as they utilize different corpora and present varying evaluation metrics. Ward and Tsukahara (2000) propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Nishimura et al. (2007) present a unimodal decision-tree approach for producing backchannels based on prosodic features. Cathcart et al. (2003) propose a unimodal model based on pause duration and trigram part-of-speech frequency. The model was constructed by identifying, from the HCRC Map Task Corpus (Anderson et al., 1991), trigrams ending with a backchannel. Fujie et al. (2004) used Hidden Markov Models to perform head nod recognition. In their paper, they combined head gesture detection with prosodic low-level features from the same person to determine strongly positive, weak positive and negative responses to yes/no type utterances.

In recent years, great research has shown the strength of latent variable models for natural language processing (Blunsom et al., 2008). One of the most relevant works is that of Eisenstein and Davis (2007), which presents a latent conditional model for fusion of multiple modalities (speech and gestures). One of the key difference of our work is that we are explicitly modeling the micro dynamics and temporal relationship between modalities.

3 Multimodal Prediction Models

Human face-to-face communication is a little like a dance, in that participants continuously adjust their behaviors based on verbal and nonverbal displays and signals. A topic of central interest in modeling such behaviors is the patterning of interlocutor actions and interactions, moment-by-moment, and one of the key challenges is identifying the patterns that best predict specific actions. Thus we are interested in developing predictive models of communication dynamics that integrate previous and current actions from all interlocutors to anticipate the most likely next actions of one or all interlocutors. Humans are good at this: they have an amazing ability to predict, at a micro-level, the actions of an interlocutor (Bavelas et al., 2000); and we know that better predictions can correlate with more empathy and better outcomes (Goldberg, 2005; Fuchs, 1987).

With turn-taking being perhaps the best-known example, we now know a fair amount about some aspects of communication dynamics, but much less about others. However, recent advances in machine learning and experimental methods, and recent findings from a variety of perspectives, including conversation analysis, social signal processing, adaptation, corpus analysis and modeling, perceptual experiments, and dialog systems-building and experimentation, mean that the time is ripe to start working towards more comprehensive predictive models.

The study of multimodal prediction models bring a new series of research challenges:

Multimodal asynchrony While speech and gestures seem to come from a single underlying mental process (McNeill, 1992), they not always happen at the same time, making it hard for earlier multimodal fusion approaches based on synchrony. A multimodal prediction model needs to be able to learn automatically the temporal relationship (and relative importance) between modalities.

Micro dynamics The dynamic between multimodal signals should be taken at a micro level since many of the interactions between speech and gesture happen at the sub-gesture level or sub-word level (Quek, 2003). Typical word-based sampling may not be sufficient and instead a higher sampling rate should be used.

Limited annotated data Given the time requirement to correctly annotate multimodal data, most multimodal datasets contain only a limited number of labeled examples. Since many machine learning algorithms rely on a large training corpus, effective training of a predictive model on multimodal datasets is challenging.

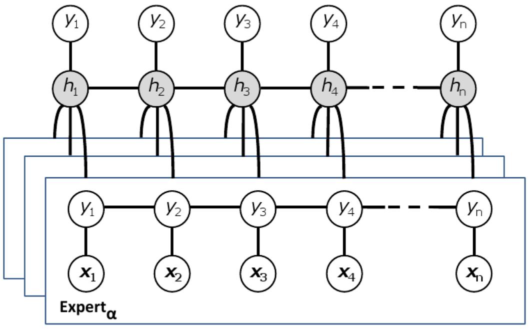


Figure 2: Latent Mixture of Discriminative Experts: a new dynamic model for multimodal fusion. In this graphical model, x_j represents the j^{th} multimodal observation, h_j is a hidden state assigned to x_j , and y_j the class label of x_j. Gray circles are latent variables. The micro dynamics and multimodal temporal relationships are automatically learned by the hidden states h_j during the learning phase.

4 Latent Mixture of Discriminative Experts

In this paper we present a multimodal fusion algorithm, called Latent Mixture of Discriminative Experts (shown in Figure 2), that addresses the three challenges discussed in the previous section. The hidden states of LMDE automatically learn the temporal asynchrony between modalities. By using a constant sample rate of 30Hz in our experiments, we can model the micro dynamics of speech and prosody (e.g., change of intonation in the middle of a word). And finally, by training separate experts for each modalities, we improve the prediction performance even with limited datasets.

The task of our LMDE model is to learn a mapping between a sequence of multimodal observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{\texttt{head-nod}, \texttt{other-gesture}\}$.

Each frame observation x_j is represented by a feature vector $\phi(x_j) \in \mathbf{R}^d$, for example, the prosodic features at each sample. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Following Morency et al. (2007), we define our LMDE model as follows:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h} \mid \mathbf{x}, \theta) \quad (1)$$

where θ is the model parameters that is to be estimated from training data.

To keep training and inference tractable, Morency et al. (2007) restrict the model to have disjoint sets of hidden states associated with each class label. Each h_j is a member of a set $\mathcal{H}_{y_j}$ of possible hidden states for the class label y_j. $\mathcal{H}$, the set of all possible hidden states, is defined to be the union of all $\mathcal{H}_y$ sets. Since sequences which have any $h_j \notin \mathcal{H}_{y_j}$ will by definition have $P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}, \theta) = 0$, latent conditional model becomes:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}: \forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h} \mid \mathbf{x}, \theta). \quad (2)$$

What differentiates our LMDE model from the original work of Morency et al. is the definition of $P(\mathbf{h}|\mathbf{x}, \theta)$:

$$P(\mathbf{h}| \, \mathbf{x}, \theta) = \frac{\exp\begin{pmatrix} \sum_l \theta_l \cdot \mathbf{T}_l(\mathbf{h}, \mathbf{x}) + \\ \sum_\alpha \theta_\alpha \cdot P_\alpha(\mathbf{y}|\mathbf{x}, \lambda_\alpha) \end{pmatrix}}{\mathcal{Z}(\mathbf{x}, \theta)}, \quad (3)$$

where $\mathcal{Z}$ is the partition function and $P_\alpha(\mathbf{y}|\mathbf{x})$ is the conditional distribution of the expert indexed by α. The expert conditional distributions are defined $P_\alpha(\mathbf{y}|\mathbf{x}, \lambda_\alpha)$ using the usual conditional random field formulation:

$$P_\alpha(\mathbf{y}| \, \mathbf{x}, \lambda_\alpha) = \frac{\exp\left(\sum_k \lambda_{\alpha,k} \cdot \mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x})\right)}{\mathcal{Z}_\alpha(\mathbf{x}, \lambda_\alpha)}, \quad (4)$$

$F_{\alpha,k}$ is defined as

$$\mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x}) = \sum_{j=1}^{m} f_{\alpha,k}(y_{j-1}, y_j, \mathbf{x}, j),$$

and each feature function $f_{\alpha,k}(y_{j-1}, y_j, \mathbf{x}, j)$ is either a state function $s_k(y_j, \mathbf{x}, j)$ or a transition function $t_k(y_{j-1}, y_j, \mathbf{x}, j)$. State functions s_k depend on a single hidden variable in the model while transition functions t_k can depend on pairs of hidden variables. $\mathbf{T}_l(\mathbf{h}, \mathbf{x})$, defined in Equation 3, is a special case, summing only over the transition feature functions $t_l(h_{l-1}, h_l, \mathbf{x}, l)$. Each expert α contains a different subset of $f_{\alpha,k}(y_{j-1}, y_j, \mathbf{x}, j)$. These feature functions are defined in Section 5.2.

4.1 Learning Model Parameters

Given a training set consisting of n labeled sequences $(\mathbf{x_i}, \mathbf{y_i})$ for $i = 1...n$, training is done in a two step process. First each expert α is trained following (Kumar and Herbert., 2003; Lafferty et al., 2001) objective function to learn the parameter λ^*_α:

$$L(\lambda_\alpha) = \sum_{i=1}^{n} \log P_\alpha(\mathbf{y}_i \mid \mathbf{x}_i, \lambda_\alpha) - \frac{1}{2\sigma^2}||\lambda_\alpha||^2 \quad (5)$$

The first term in Eq. 5 is the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\lambda_\alpha) \sim \exp\left(\frac{1}{2\sigma^2}||\lambda_\alpha||^2\right)$.

Then the marginal probabilities $P_\alpha(y_j = a \mid \mathbf{y}, \mathbf{x}, \lambda^*_\alpha)$, are computed using belief propagation and used as input for Equation 3. The optimal parameter θ^* was learned using the log-likelyhood of the conditional probability defined in Equation 2 (i.e., no regularization).

4.2 Inference

For testing, given a new test sequence $\mathbf{x}$, we want to estimate the most probable sequence of labels $\mathbf{y}^*$ that maximizes our LMDE model:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{\mathbf{h}: \forall h_i \in \mathcal{H}_{y_i}} P(\mathbf{h} \mid \mathbf{x}, \theta^*) \quad (6)$$

5 Experimental Setup

We evaluate our Latent Mixture of Discriminative Experts on the multimodal task of predicting listener nonverbal backchannel (i.e., head nods). Backchannel feedback (the nods and paraverbals such as "uh-hu" and "mm-hmm" that listeners produce as some is speaking) has received considerable interest due to its pervasiveness across languages and conversational contexts.

5.1 Dataset

We are using the RAPPORT dataset from (Maatman et al., 2005), which contains 47 dyadic interactions between a speaker and a listener. Data is drawn from a study of face-to-face narrative discourse ("quasi-monologic" storytelling). In this dataset, participants in groups of two were told they were participating in a study to evaluate a communicative technology. Subjects were randomly assigned the role of speaker and listener. The speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants sat approximately 8 feet apart. Video sequences were manually annotated to determine the ground truth head nod labels. A total of 587 head nods occured over all video sequences.

5.2 Multimodal Features

This section describes the different multimodal features used to create our five experts.

PROSODY Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody (Nishimura et al., 2007; Ward and Tsukahara, 2000; Cathcart et al., 2003). For example, Ward and Tsukahara (2000) show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. Listener feedback often follows speaker pauses or filled pauses such as "um" (see (Cathcart et al., 2003)). Using openSMILE (Eyben et al., 2009) toolbox, we extract the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara: downslopes in pitch continuing for at least 40ms, regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness), drop or rise in energy of speech (i.e., energy edge), Fast drop or rise in energy of speech (i.e., energy fast edge), vowel volume (i.e., vowels are usually spoken softer) and Pause in speech (i.e., no speech).

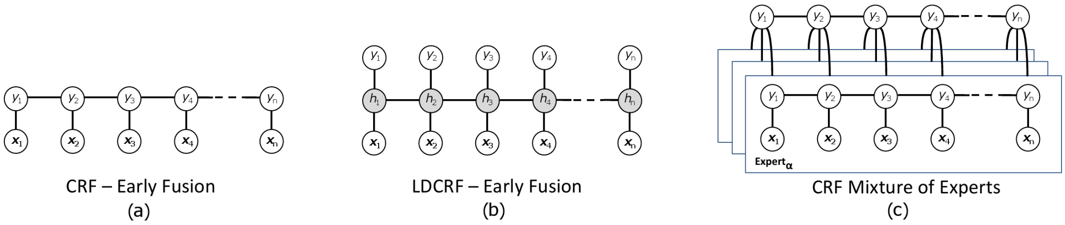
VISUAL GESTURES Gestures performed by the speaker are often correlated with listener feedback (Burgoon et al., 1995). Eye gaze, in particular, has often been implicated as eliciting listener feedback. Thus, we manually annotate the following contextual feature: speaker looking at the listener.

LEXICAL Some studies have suggested an association between lexical features and listener feedback (Cathcart et al., 2003). Using the transcriptions, we included all individual words (i.e., unigrams) spoken by the speaker during the interactions.

SYNTACTIC STRUCTURE Finally, we attempt to capture syntactic information that may provide relevant cues by extracting four types of features from a syntactic dependency structure corresponding to the utterance. The syntactic structure is produced automatically using a CRF part-of-speech (POS) tagger and a data-driven left-to-right shift-reduce dependency parser (Sagae and Tsujii, 2007), both trained on POS tags and dependency trees extracted from the Switchboard section of the Penn Treebank (Marcus et al., 1994), converted to dependency trees using the Penn2Malt tool[1]. The four syntactic features are:

- Part-of-speech tags for each word (e.g. noun, verb, etc.), taken from the output of the POS tagger
- Grammatical function for each word (e.g. subject, object, etc.), taken directly from the dependency labels produced by the parser
- Part-of-speech of the syntactic head of each word, taken from the dependency links produced by the parser
- Distance and direction from each word to its syntactic head, computed from the dependency links produced by the parser

[1] http://w3.msi.vxu.se/ nivre/research/Penn2Malt.html



CRF – Early Fusion
(a)

LDCRF – Early Fusion
(b)

CRF Mixture of Experts
(c)

Figure 3: Baseline Models: **a)** Conditional Random Fields (CRF), **b)** Latent Dynamic Conditional Random Fields(LDCRF), **c)** CRF Mixture of Experts (no latent variable)

Although our current method for extracting these features requires that the entire utterance be available for processing, this provides us with a first step towards integrating information about syntactic structure in multimodal prediction models. Many of these features could in principle be computed incrementally with only a slight degradation in accuracy, with the exception of features that require dependency links where a word's syntactic head is to the right of the word itself. We leave an investigation that examines only syntactic features that can be produced incrementally in real time as future work.

5.3 Baseline Models

INDIVIDUAL EXPERTS Our first baseline model consists of a set of CRF chain models, each trained with different set of multimodel features (as described in the previous section). In other words, only visual, prosodic, lexical or syntactic features are used to train a single CRF expert. In one CRF chain model, each gesture class corresponds to a state label. (See Figure 3a).

MULTIMODAL CLASSIFIERS (EARLY FUSION) Our second baseline consists of two models: CRF and LDCRF (Morency et al., 2007). To train these models, we concatenate all multimodal features (lexical, syntactic, prosodic and visual) in one input vector. Graphical representation of these baseline models are given in Figure 3.

CRF MIXTURE OF EXPERTS To show the importance of latent variable in our LMDE model, we trained a CRF-based mixture of discriminative experts. This model is similar to the Logarithmic Opinion Pool (LOP) CRF suggested by Smith et al. (2005). The training is performed in two steps. A graphical representation of a CRF Mixture of experts is given in the last graph of Figure 3.

5.4 Methodology

We performed held-out testing by randomly selecting a subset of 11 interactions (out of 47) for the test set. The training set contains the remaining 36 dyadic interactions. All models in this paper were evaluated with the same training and test sets. Validation of all model parameters (regularization term and number of hidden states) was performed using a 3-fold cross-validation strategy on the training set. The regularization term was validated with values $10^k, k = -1..3$. Three different number of hidden states were tested for the LMDE models: 2, 3 and 4.

The performance is measured by using the F-measure. This is the weighted harmonic mean of precision and recall. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model. We use the same weight for both precision and recall, so-called F_1. During validation we find all the peaks (i.e., local maxima) from the marginal probabilities. These backchannel hypotheses are filtered using the optimal threshold from the validation set. A backchannel (i.e., head nod) is predicted correctly if a peak happens during an actual listener backchannel with high enough probability. The same evaluation measurement is applied to all models.

The training of all CRFs and LDCRFs were done using the hCRF library[2]. The LMDE model was implemented in Matlab[3] based on the hCRF

[2] http://sourceforge.net/projects/hrcf/

[3] The source code is available at: http://projects.ict.usc.edu/multicomp/.

Table 1: Comparison of individual experts with our Latent Mixture of Discriminative Experts (LMDE).

Expert	Precision	Recall	f1
Lexical	0.1647	0.3305	0.2198
Prosody	0.1396	0.9112	0.2421
Syntactic	0.1833	0.4663	0.2632
POS	0.1935	0.4514	0.2709
Eye Gaze	0.1573	0.1741	0.1653
LMDE	0.2295	0.5677	**0.3268**

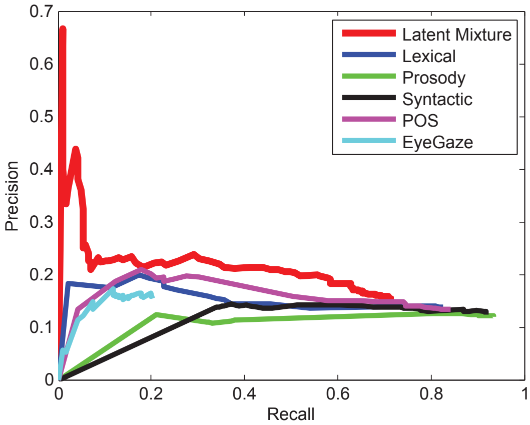


Figure 4: Comparison of individual experts with our LMDE model.

library.

6 Results and Discussion

In this section we present the results of our empirical evaluation designed to test the three main characteristics of the LMDE model: (1) integration of multiple sources of information, (2) late fusion approach and (3) latent variable which models the hidden dynamic between experts. We also present an analysis of the output probabilities from the LMDE model and individual experts.

Individual Experts We trained one individual expert for each feature types: visual, prosodic, lexical and syntactic features (both part-of speech and syntactic structure). Precision, recall and F_1 values for each individual expert and our LMDE model are shown in Table 1 and Figure 4.

Pairwise two-tailed t-test comparison between our LMDE model and individual experts shows a significant difference for Lexical, Prosody, Syntactic and Eye gaze, with respective p-values of 0.0037, 0.0379, 0.0400 and 0.0233. Even though some experts may not perform well individually (e.g., eye gaze), they can bring important information once merged with others. Table 1 shows that our LMDE model was able to take advantage of the complementary information from each expert.

Table 2: Comparison of our Latent Mixture of Discriminative Experts (LMDE) with two early fusion technique (CRF vs LDCRF) and the CRF Mixture of Experts (Smith et al., 2005).

model	Precision	Recall	f1
LMDE	0.2295	0.5677	**0.3268**
Early CRF	0.13958	0.9245	0.2425
Early LDCRF	0.1826	0.2484	0.2105
Mixture CRF	0.1502	0.2712	0.1934

Late Fusion We compare our approach with two early fusion models: CRF and Latent-dynamic CRF (see Figure 3). Table 2 summarizes the results. The CRF model learns direct weights between input features and the gesture labels. The LDCRF is able to model more complex dynamics between input features with the latent variable. We can see that our LMDE model outperforms both early fusion approaches because of its late fusion approach. Pairwise two-tailed t-test analysis gives p-values of 0.0481 and 0.0748, for CRF and LDCRF respectively.

Latent Variable The CRF Mixture of Experts (2005) directly merges the expert outputs while our model uses a latent variable to model the hidden dynamic between experts (see Figure 3). Table 2 summarizes the results. Pairwise two-tailed t-test comparison between these two models shows a significant difference with a p-value of 0.0062. This result is important since it shows that our LMDE model does learn the hidden interaction between experts.

Model analysis To understand the multimodal integration which happens at the latent variable level in our LMDE model, Figure 5 shows the output probabilities for all five individual experts as well as our model. The strength of the latent variable is to enable different weigting

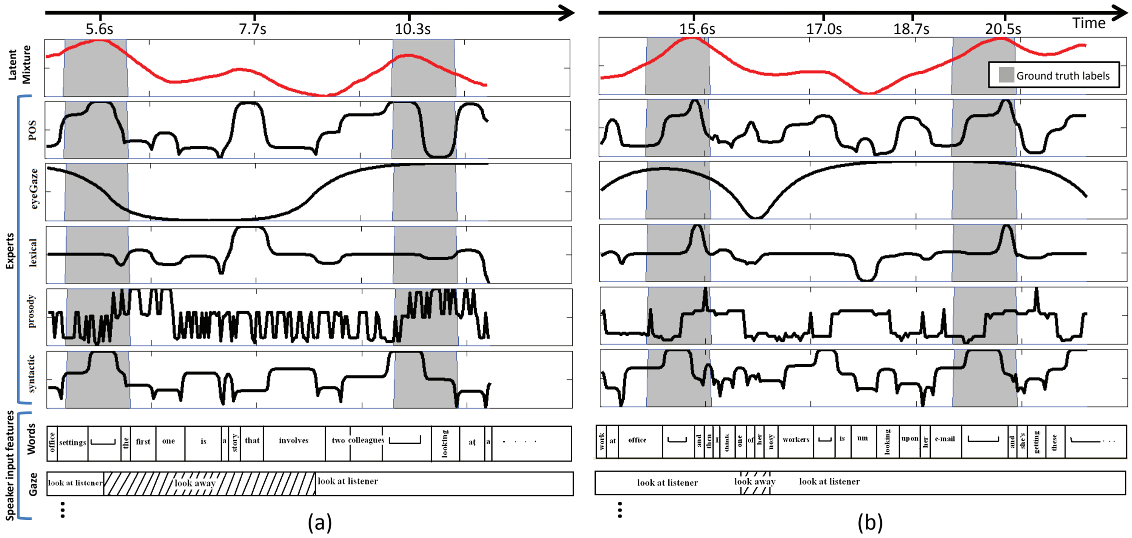


Figure 5: Output probabilities from LMDE and individual experts for two different sub-sequences. The gray areas in the graph corresponds to ground truth backchannel feedbacks of the listener.

of the experts at different point in time.

By analyzing the sequence (a), we observe that both the POS and Syntactic experts learned that when no words are present (i.e., pause) there is a high likelihood of backchennel feedback from the listener (shown at 5.6s and 10.3s). These two experts are highly weighted (by one of the hidden state) during this part of the sequence. Also, both the Lexical and POS experts learned that the word "‘that"’ (and its part-of-speech) are important but since the speaker is not looking at the listener when saying it, the output from LMDE model is low (see Figure 5, Sequence (a), 7.7s).

By analyzing sequence (b), we see that the Lexical and POS experts learned the importance of the "‘and"’ at 15.6s and 20.5s. More importantly, we can see at 17.0s and 18.7s that the influence of the POS and Syntactic experts have been reduced in the LMDE output probability. This difference of weighting shows that a different hidden state is active during Sequence (b).

7 Conclusion

In this paper we introduced a new model called Latent Mixture of Discriminative Experts (LMDE) for learning predictive models of human communication behaviors. Many of the interactions between speech and gesture happen at the sub-gesture or sub-word level. LMDE learns automatically the temporal relationship between different modalities. Since, we train separate experts for each modality, LMDE is capable of improving the prediction performance even with limited datasets.

We evaluated our model on the task of nonverbal feedback prediction (e.g., head nod). Our experiments confirm the importance of combining the four types of multimodal features: lexical, syntactic structure, eye gaze, and prosody. LMDE is a generic model that can be applied to a wide range of problems. As future work, we are planning to test our model on dialog act classification and multimodal behavior recognition tasks.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Anderson, H., M. Bader, E.G. Bard, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo,

H. Thompson, and R. Weinert. 1991. The mcrc map task corpus. *Language and Speech*, 34(4):351–366.

Bavelas, J.B., L. Coates, and T. Johnson. 2000. Listeners as co-narrators. *JPSP*, 79(6):941–952.

Blunsom, P., T. Cohn, and M. Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *ACL: HLT*, pages 200–208.

Burgoon, Judee K., Lesa A. Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, Cambridge.

Cassell, J. and M. Stone. 1999. Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems. In *AAAI*.

Cathcart, N., Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *EACL*, pages 51–58.

Eisenstein, J., R. Barzilay, and R. Davis. 2008. Gestural cohesion for topic segmentation. In *ACL: HLT*, pages 852–860.

Eisentein, J. and R. Davis. 2007. Conditional modality fusion for coreference. In *ACL*, pages 352–359.

Eyben, Florian, Martin Wöllmer, and Björn Schuller. 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *ACII*, pages 576–581.

Frampton, M., J. Huang, T. Bui, and S. Peters. 2009. Real-time decision detection in multi-party dialogue. In *EMNLP*, pages 1133–1141.

Fuchs, D. 1987. Examiner familiarity effects on test performance: implications for training and practice. *Topics in Early Childhood Special Education*, 7:90–104.

Fujie, Shinya, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. 2004. A conversation robot using head gesture recognition as para-linguistic information. In *RO-MAN*, pages 159–164.

Goldberg, S.B. 2005. The secrets of successful mediators. *Negotiation Journal*, 21(3):365–376.

Gravano, A., S. Benus, H. Chavez, J. Hirschberg, and L. Wilcox. 2007. On the role of context and prosody in the interpretation and 'okay'. In *ACL*, pages 800–807.

Heylen, D. and R. op den Akker. 2007. Computing backchannel distributions in multi-party conversations. In *ACL:EmbodiedNLP*, pages 17–24.

Johnston, M. 1998. Multimodal language processing. In *ICSLP*.

Jovanovic, N., R. op den Akker, and A. Nijholt. 2006. Adressee identification in face-to-face meetings. In *EACL*.

Jurafsky, D., E. Shriberg, B. Fox, and T. Curl. 1998. Lexical, prosodic and syntactic cures for dialog acts. In *Workshop on Discourse Relations*, pages 114–120.

Kendon, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kumar, S. and M. Herbert. 2003. Discriminative random fields: A framework for contextual interaction in classification. In *ICCV*.

Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *ICML*.

Maatman, M., J. Gratch, and S. Marsella. 2005. Natural behavior of a listening agent. In *IVA*.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: annotating predicate argument structure. In *ACL:HLT*, pages 114–119.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press.

Moore, P.-Y. Hsueh J. 2007. What decisions have you made: Automatic decision detection in conversational speech. In *NAACL-HLT*, pages 25–32.

Morency, Louis-Philippe, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*.

Murray, G. and G. Carenini. 2009. Predicting subjectivity in multimodal conversations. In *EMNLP*, pages 1348–1357.

Nakano, Reinstein, Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *ACL*.

Nakano, Y., K. Murata, M. Enomoto, Y. Arimoto, Y. Asa, and H. Sagawa. 2007. Predicting evidence of understanding by monitoring user's task manipulation in multimodal conversations. In *ACL*, pages 121–124.

Nishimura, Ryota, Norihide Kitaoka, and Seiichi Nakagawa. 2007. A spoken dialog system for chat-like conversations considering response timing. *LNCS*, 4629:599–606.

Oviatt, S. 1999. Ten myths of multimodal interaction. *Communications of the ACM*.

Quek, F. 2003. The catchment feature model for multimodal language analysis. In *ICCV*.

Sagae, Kenji and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *ACL*, pages 1044–1050.

Smith, A., T. Cohn, and M. Osborne. 2005. Logarithmic opinion pools for conditional random fields. In *ACL*, pages 18–25.

Ward, N. and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207.