INTEGRATING THUMBNAIL FEATURES FOR SPEECH RECOGNITION USING CONDITIONAL EXPONENTIAL MODELS

Hua Yu and Alex Waibel

Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA 15213
{hyu,ahw}@cs.cmu.edu

ABSTRACT

We describe a novel approach for modeling segmental information in speech recognition, through the use of thumbnail features. By taking into account dependencies at the segmental level, thumbnail features are more resistant to changes in speaking rates and other factors. While the traditional acoustic features are fixed for every utterance, one set of thumbnail features is computed for each hypothesis, which may violate the traditional scoring paradigm. To this end, we introduce a conditional exponential modeling framework. It allows better integration of various knowledge sources in a discriminative fashion. We present preliminary experiments on the Switchboard task.

1. INTRODUCTION

Speech recognition using Hidden Markov Models (HMMs) has achieved remarkable success over the years. However, HMM has one well noted weakness: the *frame independence assumption*. Each acoustic frame is independent of any other, given the state sequence. In other words, HMM is performing mostly pattern matching on a frame by frame basis, ignoring any higher level structures.

Many proposals have been made to rectify the situation, such as segmental models [1], parallel path HMMs [2], buried HMMs [3], Gaussian transition models [4], and so on.

We propose a novel *thumbnail* feature, where we directly compute a feature vector for each segment and model it as a whole. A segment can be either a phone, a syllable, or even a word. The feature vector can be thought of as a thumbnail image of the entire phone/syllable/word. In other words, we are now modeling each segment as one unit, rather than a sequence of frames. This has been shown to be quite useful in handwriting recognition.

Since thumbnail features depend on the segmentation of speech, they cannot be pre-computed as in the traditional acoustic feature analysis. They have to be computed in a hypothesis-driven way. Comparing scores between hypotheses also becomes an issue. This is discussed in Section 2.

An even more important issue is how to integrate this new knowledge source with the conventional frame-based acoustic model, as well as other knowledge sources. Motivated by recent works in the machine learning field, especially the work on conditional random fields by Lafferty [5], we propose to use conditional exponential models as the overall framework. This framework is nicely suited for our task. It allows the use of arbitrary input features (such as segmental features), and is a discriminative model by design. Section 3 gives the details.

Section 4 describes preliminary experimental results on the Switchboard task.

2. THUMBNAIL FEATURES

Conventional speech analysis (front-end) provides a sequence of frames as input to the acoustic model. Due to the aforementioned frame independence assumption, each triphone model can only have a very "short-sighted" view of the data, one frame at a time. Dependencies between frames are completely ignored. This is clearly a problem.

Thumbnail features attempt to capture the global characteristics of speech segments, similar to thumbnails of images. In this paper, we choose to sub-sample the sequence of frames of a segment to arrive at a feature vector of certain fixed dimensionality. We note, however, that the vector needs not to have fixed dimensionality and it can be computed in a more intelligent fashion.

Modeling a segment as a whole has many advantages. First, it captures segmental level structures. Second, it is more robust with respect to changes in speaking rates. For example, certain parts of a phoneme may be heavily reduced in fast speech. A frame-by-frame matching would perform poorly in this case. As the overall shape of the phoneme is better preserved, we expect the thumbnail feature to perform better. Greenberg argued that even entire phones could be deleted in conversational speech [6]. It would then be more appropriate to use thumbnail features on the syllable level.

2.1. Intelligent Dynamic Features

Our particular implementation of thumbnail features can be viewed as an intelligent extension of the dynamic features.

Dynamic features, also known as the delta and double-delta features [7], are well known to be very effective in improving recognition accuracy. Although from a generative point of view, it is completely wrong to generate from an HMM a sequence of feature vectors that contain the delta information. Such a sequence of feature vectors may not correspond to a real cepstral sequence. Nonetheless, dynamic features help to alleviate the frame independence assumption and contribute to a more discriminative model [8].

The delta and double-delta features are typically computed from adjacent frames. If we instead compute them from frames at carefully chosen locations in a segment, the set of frames can be viewed as a thumbnail sketch of the segment.

2.2. Hypothesis Driven Features

There is a salient distinction between thumbnail features and conventional features. Unlike conventional features (front-ends),

thumbnail features cannot be computed beforehand and fixed during recognition. To compute thumbnail features, we need segmentation information, which is different for each hypothesis. In other words, each hypothesis has its own observation space.

We therefore face a "chicken-and-egg" problem. The features cannot be computed without a segmentation; but to find a reasonable segmentation (or the hypothesis that implies the segmentation), we need to begin with some acoustic feature. To avoid the problem of simultaneously searching for both the feature and the hypothesis, we adopt a two stage approach. A conventional system is used to find a set of hypotheses and their segmentation. Then, thumbnail features are computed and hypotheses are rescored using a second set of thumbnail models.

It is worth noting that this scenario, i.e. dependencies of a state on arbitrary input frames, is one of the hallmarks of exponential models, which we introduce in Section 3.

2.3. Score Integration Issues

In the conventional speech recognition framework,

$$p(\boldsymbol{h}|\boldsymbol{o}) \propto p(\boldsymbol{o}|\boldsymbol{h})p(\boldsymbol{h})$$

where $\boldsymbol{h}$ is a hypothesis, $\boldsymbol{o} = (o_1, o_2, \cdots, o_T)$ is the observation sequence, $p(\boldsymbol{o}|\boldsymbol{h})$ gives the acoustic score, $p(\boldsymbol{h})$ gives the language model score. $p(\boldsymbol{o}|\boldsymbol{h}) = \prod_t p(o_t|\boldsymbol{h})$ under the frame independence assumption.

Since each hypothesis has its own observation space under segmental modeling, comparing scores between them can be a problem. Hypotheses with fewer segments will have fewer probability terms and better scores in general.

One solution is to enforce the same number of segments for each hypothesis by computing a thumbnail feature for each frame. This makes the score comparable between different hypotheses.

Furthermore, since segmental models provide complementary information to the traditional acoustic model, we would want to combine segmental scores with frame-based acoustic scores. Again, it is not clear how to combine them optimally. This gives a second motivation for the conditional exponential model, where all the knowledge sources can be viewed as *features* and can be combined in a discriminative framework.

3. CONDITIONAL EXPONENTIAL FRAMEWORK

3.1. Background

There has been a growing interest in going beyond the HMM framework. A desirable new framework should be discriminative, and can make use of arbitrary features which, for example, captures long-range dependencies of the observations.

- Conditional model: HMMs assign a probability to a frame sequence given a word sequence: $p(\boldsymbol{o}|\boldsymbol{h})$. As a generative model, many parameters are needed to cover the entire acoustic space. But during recognition, the observation is given anyway. A conditional model, $p(\boldsymbol{h}|\boldsymbol{o})$, on the other hand, does not need to spend parameters on covering the observation space. This leads to a more compact model.
- Arbitrary features: HMMs depend on the conditional independence assumption to achieve tractability. Any dependencies between frames, such as segmental features, pose a problem for HMMs. To handle multiple mutually dependent features, exponential models are a natural choice.

In recent AI research, there are several proposals that achieves both: Maximum Entropy Markov Models (MEMM) [9], and the more general Conditional Random Fields (CRF) [5]. Roughly speaking, a *conditional random field* is an exponential model on the whole sequence level, which models the conditional distribution $p(\boldsymbol{h}|\boldsymbol{o})$. The state dependency in a CRF follows a graph structure. In practice, most CRFs assume a linear chain graph, which is similar to the state dependencies in an HMM.

If we further drops the graph dependency assumptions, we arrive at a conditional exponential model for the whole sequence:

$$p(\boldsymbol{h}|\boldsymbol{o}) = \frac{1}{Z(\boldsymbol{o})} \exp\left(\sum_i \lambda_i f_i(\boldsymbol{o}, \boldsymbol{h})\right) \quad (1)$$

where $\boldsymbol{h}$ is a hypothesized sentence, $\boldsymbol{o}$ a frame sequence, $f_i(\boldsymbol{o}, \boldsymbol{h})$ is a feature function, $Z(\boldsymbol{o})$ the normalization factor.

After being proposed for natural language processing [10], exponential models have been used in many tasks, such as shallow parsing, statistical translation [11]. It has also been used to enhance a n-gram language model with higher level features [12]. Recently, Likhodedov et al. extended MEMM to phone recognition [13], and Macherey et al. applied exponential models to acoustic modeling for connected digit recognition [14]. But for most of the speech community, this is still relatively new.

We will first give a brief introduction to exponential models (see [10] for details), then move on to conditional exponential models, and discuss how they can be applied to speech recognition.

3.2. Exponential Model: a Feature-Rich Model

Exponential models arise naturally in a maximum entropy setting. In most statistical modeling scenarios, we don't necessarily know the right model family. But we do know a set of statistics that captures the behavior of the underlying random process, and we want our model to match these statistics. More formally, to model the distribution of random variable x, we first define a set of *feature functions* $\{f_i(x)\}$. $f_i(x)$ can be any binary or real valued functions. We would like

$$\tilde{p}(f_i) = p(f_i), \forall i$$

where $\tilde{p}(f_i)$ is the expectation of f_i on the training data, $p(f_i)$ is the expectation of f_i with respect to our model. When x is discrete, $p(f_i) \equiv \sum_x p(x) f_i(x)$. For example, by taking the number of words in a hypothesis to be a feature, we introduce a constraint on the average sentence length.

According to the maximum entropy principle, among all distributions that satisfy the set of constraints, we should choose the one that assumes least about the unknowns, i.e. the one with the highest entropy. It can be shown [10] that the maximum entropy solution takes an exponential form:

$$p(x) = \frac{1}{Z} \exp\left(\sum_i \lambda_i f_i(x)\right) \quad (2)$$

where λ_i's are parameters to be estimated (one per feature), Z is a normalizing constant. Berger et al. proved that the maximum entropy solution also achieves maximum likelihood among distributions of form (2).

3.3. Conditional Exponential Models

We now extend the maximum entropy principle to model the conditional distributions $p(y|x)$, where y is the class label of x. The feature function $f_i(x, y)$ is now defined on the (x, y) pair. The constraints are

$$\tilde{p}(f_i) = p(f_i), \forall i$$

where

$$\tilde{p}(f_i) \equiv \sum_{x,y} \tilde{p}(x, y) f(x, y)$$

$$p(f_i) \equiv \sum_{x,y} \tilde{p}(x) p(y|x) f(x, y)$$

where $\tilde{p}(x, y)$ is the empirical joint distribution, $\tilde{p}(x)$ is the corresponding marginal distribution.

The maximum entropy solution takes the following form:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x, y)\right) \quad (3)$$

Note the normalization term is now dependent on x:

$$Z(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x, y)\right)$$

3.4. Conditional Exponential Model for Speech Recognition

Let $\boldsymbol{h}$ be a word sequence, o be a frame sequence, we can apply Equation 3 to speech recognition:

$$p(\boldsymbol{h}|\boldsymbol{o}) = \frac{1}{Z(\boldsymbol{o})} \exp\left(\sum_i \lambda_i f_i(\boldsymbol{o}, \boldsymbol{h})\right)$$

Since it models the probability of any hypothesis given an observation, it is a discriminative model by design. And there is no need to expend parameters on covering the observation space. It can focus all the modeling power on decision boundaries.

Since there is no feature independence requirements, we can use *arbitrary* features: local frame level features, segmental features, or sentence level features, as long as they contribute to a more discriminative model.

The model is based on a set of feature functions. They can be any computable quantities. We can even re-formulate the conventional ASR score computation in this new model. Say we have a conventional acoustic model $\mathcal{AM}$ that computes the acoustic score $p_{\mathcal{AM}}(\boldsymbol{o}|\boldsymbol{h})$, and a language model that computes $p_{\mathcal{LM}}(\boldsymbol{h})$. Let

$$\begin{aligned} f_1(\boldsymbol{o}, \boldsymbol{h}) &= \log p_{\mathcal{AM}}(\boldsymbol{o}|\boldsymbol{h}) \\ f_2(\boldsymbol{o}, \boldsymbol{h}) &= \log p_{\mathcal{LM}}(\boldsymbol{h}) \\ f_3(\boldsymbol{o}, \boldsymbol{h}) &= \#(\text{words in } \boldsymbol{h}) \end{aligned}$$

then

$$\begin{aligned} \log p(\boldsymbol{h}|\boldsymbol{o}) &= -\log Z(\boldsymbol{o}) + \lambda_1 \log p_{\mathcal{AM}}(\boldsymbol{o}|\boldsymbol{h}) \\ &\quad + \lambda_2 \log p_{\mathcal{LM}}(\boldsymbol{h}) + \lambda_3 \#(\text{words in } \boldsymbol{h}) \quad (4) \end{aligned}$$

This is clearly equivalent to the conventional ASR score combination formula. λ_1 and λ_2 together decide the relative weight between the acoustic model and the language model, λ_3 is the word insertion penalty. Rather than being just "fudge factors", they can now be estimated in a discriminative fashion.

We can easily add more knowledge sources, such as a second acoustic model, segmental models and language models as extra features, and achieve discriminative model combination.

Beyerlein [15] describes a similar model combination form, using a different objective function. But conditional exponential models can do far more than just model combination. We can apply it directly to acoustic modeling, like [14], by defining appropriate features.

3.5. Model Training

As mentioned before, maximum likelihood solution and maximum entropy solution are the same for exponential models. Thus, unlike in HMMs, discrimination can be achieved by ML training. Better yet, the optimization problem is convex [16]. Hence, we can use any iterative procedure, such as iterative scaling, conjugate gradient descent or quasi-Newton methods, to find the global maximum. Let $\Lambda = \{\lambda_i\}$ be the current estimate of model parameters, $L_{\tilde{p}}(\Lambda)$ be the likelihood function to be optimized, it can be shown that:

$$\begin{aligned} L_{\tilde{p}}(\Lambda) &= \sum_{(\boldsymbol{O}, \boldsymbol{S})} \tilde{p}(\boldsymbol{o}, \boldsymbol{s}) \log p_\Lambda(\boldsymbol{s}|\boldsymbol{o}) \\ \frac{\partial L_{\tilde{p}}(\Lambda)}{\partial \lambda_i} &= \tilde{p}(f_i) - p_\Lambda(f_i) \end{aligned}$$

where $\boldsymbol{s}$ is the transcript for $\boldsymbol{o}$.

The difficulty, however, is to compute the normalization term

$$Z(\boldsymbol{o}) = \sum_{\boldsymbol{h}} \exp\left(\sum_i \lambda_i f_i(\boldsymbol{o}, \boldsymbol{h})\right)$$

The first summation is over all possible word sequences, which is a daunting task. Similar to other discriminative training methods, such as maximum mutual information training [17], we can approximate $Z(\boldsymbol{o})$ by restricting the summation to sentences in a lattice or an N-Best list [18]. Therefore, a decoding pass needs to be carried out on the training data. In the case when reference transcript is not in the lattice, we need to add/merge it into the lattice.

4. EXPERIMENTS

Experiments are carried out on the Switchboard task. The test set is a 1 hour subset of the 2001 Hub5e evaluation set. A 66 hour subset of the SWB-I corpus is used for training. The baseline system uses vocal tract length normalization, cluster-based cepstral mean normalization, and an 11-frame context window for delta and double-delta. Linear discriminant analysis (LDA) is applied to reduce feature dimensionality to 42. There are a total of 86k Gaussians in the triphone-clustered acoustic models. For testing, we use a 15k vocabulary and a trigram language model trained on SWB and CallHome. The baseline system has a word error rate (WER) of 34.5%.

The baseline system is used to produce N-Best lists (N=800) and the associated segmentations. With the segmentation, we can compute thumbnail features for each hypothesis and rescore the N-Best lists.

In the first experiment, we compute thumbnail features on a frame-by-frame basis to avoid score normalization issues. For

each frame, we first calculate its relative position in a segment (either sub-phone segment or phone segment). Frames at the same relative position in neighboring segments are extracted, and stacked together to form a super-vector, which is then projected onto a 42-dimensional subspace using LDA. The LDA matrix is estimated from the training data, in a way similar to computing the LDA matrix for the baseline system. An acoustic model set can then be trained using the projected thumbnail features. We experimented with thumbnail features at two different levels: sub-phone level and phone level. At the sub-phone level, we get a 0.3% improvement (34.5% $\rightarrow$ 34.2%). At the phone level, WER is reduced significantly to 33.7%. The best performance is achieved using thumbnail features computed from 5 adjacent phone segments.

In the second experiment, we use the conditional exponential model for score combination (Equation 4). Rather than combining the overall acoustic/language/segmental model scores, we believe it makes more sense to combine the *average* model score. The overall score tends to favor short hypotheses, because it is roughly proportional to the length of the hypothesis, or the number of segments. Let

$$f_1(\boldsymbol{o},\boldsymbol{h}) = \frac{\log p_{\mathcal{AM}}(\boldsymbol{o}|\boldsymbol{h})}{\#(\text{frames in } \boldsymbol{o})} \qquad f_2(\boldsymbol{o},\boldsymbol{h}) = \frac{\log p_{\mathcal{LM}}(\boldsymbol{h})}{\#(\text{words in } \boldsymbol{h})}$$

$$f_3(\boldsymbol{o},\boldsymbol{h}) = \#(\text{words in } \boldsymbol{h}) \qquad f_4(\boldsymbol{o},\boldsymbol{h}) = \frac{\log p_{\mathcal{SM}}(\boldsymbol{o}|\boldsymbol{h})}{\#(\text{segments in } \boldsymbol{h})}$$

where $\mathcal{SM}$ denotes the segmental model, the new model should be more fair to long hypotheses. Notice here, instead of computing segmental scores on a frame-by-frame basis, we compute a single segmental score for each segment.

Using just the segmental model with phone level thumbnail features, the language model and sentence length feature, we get 33.6% by rescoring the N-Best lists. Adding the baseline acoustic model as an additional feature, WER becomes 33.4%. Hence, our unconventional scoring formula gives slightly better results compared to the original scoring paradigm. We expect that larger improvements can be obtained by adding more features.

5. CONCLUSIONS

We introduced a novel thumbnail feature for segmental modeling, and a conditional exponential modeling framework to integrate multiple knowledge sources. We showed encouraging results (34.5%$\rightarrow$33.4%) in preliminary experiments. The new framework is discriminative in nature, and can take advantages of arbitrary features. With the increase in computing power, it has the potential to supersede the HMM framework. We are currently working on replacing N-Best lists with lattices, as well as adding more interesting features to the exponential model.

6. ACKNOWLEDGMENTS

We thank Rong Yan, Yan Liu, Jian Zhang, Xiaojin Zhu and John Lafferty for insightful discussions.

7. REFERENCES

[1] M. Ostendorf, V. Digilakis, and O. Kimball, "From hmms to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, 1996.

[2] R. Iyer, H. Gish, M. Siu, G. Zavaliagkos, and S. Matsoukas, "Hidden markov models for trajectory modeling," in *Proc. ICSLP*, 1998.

[3] J. Bilmes, "Buried markov models for speech recognition," in *Proc. ICASSP*, 1999.

[4] H. Yu and T. Schultz, "Implicit trajectory modeling through gaussian transition models for speech recognition," in *Proceedings of the Human Language Technology Conference (HLT)*, 2003.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning*, 2002.

[6] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, 1998.

[7] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans ASSP*, vol. 34, no. 1, pp. 52–59, 1986.

[8] H. Yu, "Phase-space representation of speech — revisiting the delta and double delta features," in *http://www.cs.cmu.edu/~ hyu/pspace.pdf*, 2003.

[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proceedings of the International Conference on Machine Learning*, 2001.

[10] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, pp. 39–71, 1996.

[11] F. Och, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[12] R. Rosenfeld, "A whole sentence maximum entropy language model," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*, 1997.

[13] A. Likhodedov and Y. Gao, "Direct models for phoneme recognition," in *Proc. ICASSP*, 2002.

[14] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Proc. EuroSpeech*, 2003.

[15] P. Beyerlein, "Discriminative model combination," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*, 1997.

[16] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 1–13, 1997.

[17] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *ISCA ITRW Automatic Speech Recognition*, 2000.

[18] R. Schwartz and S. Austin, "A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses," in *Proc. ICASSP*, 1993.