CLASSIFICATION AND RECOGNITION WITH DIRECT SEGMENT MODELS

Geoffrey Zweig

Microsoft Research
gzweig@microsoft.com

ABSTRACT

Segment based direct models have recently been used to improve the output of existing state-of-the-art speech recognizers. To date, however, they have relied on an existing HMM system to provide segment boundaries. This paper takes initial steps at using these models on their own, first by developing a segment-based maximum entropy phone classifier, and then by utilizing the features in a segmental conditional random field for recognition. To produce a feature representation that is independent of segment length, we utilize a set of ngram features based on vector-quantized representations of the acoustic input. We find that the models are able to integrate information at different granularities and from different streams. Contextual information from around the segment boundaries is particularly important. We obtain competitive results for TIMIT phone classification, and present initial recognition results.

Index Terms— Segmental Conditional Random Fields, Maximum Entropy, Speech Recognition

1. INTRODUCTION

Recently, a number of direct segmental models have been proposed. Segmental Conditional Random Fields (SCRFs) [1], Conditional Augmented Models [2] and Structured SVMs [3] all perform a segment-level analysis of an utterance, using features which are fundamentally different from those available in a frame-wise analysis. Such segment models have two basic advantages. First, they enable new classes of features to be used, where segment boundaries are an integral part of the feature definition, and frame-wise conditional independence assumptions are no longer present. Examples of these features are segment length [4], template matching distances [5, 6, 7], and Fisher Kernel scores [2, 3]. The second advantage is that their log-linear form allows for the coherent integration of many different types of features. For example, [8] integrates binary, ordinal and real feature values.

To date, these models have been used to improve on the output of existing state-of-the-art recognizers [1, 2, 3, 8], but never independently of an existing HMM system, which has been used to provide a lattice of potential segmentations. In this paper, we take some initial steps at using an unaided SCRF for recognition, first by developing and testing a set of features for maximum entropy phone classification, and then by using them as the basis for SCRF recognition. Due to the model structure, the key problem that must be addressed is how to represent a variable length segment with a fixed length feature vector. In [2, 3], this is done by using the likelihood and Fisher Kernel scores of a generative model applied to the segment. Other approaches have been based on frame-averaging [9] and sub-sampling [10]. In this work, we use pattern matching within a discrete vector-quantized representation of the acoustic signal. Specifically, the feature vector consists of indicator variables signaling the presence or absence of VQ ngram patterns within the segment. Additionally, we use position-dependent patterns (such as the presence of an ngram prefix) which explicitly refer to the segment boundaries. We believe the use of continuous features may result in improvement; however, we find a discrete VQ representation to provide reasonable results with low computational requirements, and robust to parameter optimization techniques, and thus suitable for developing the framework.

2. MODELS STUDIED

2.1. Maximum Entropy for Classification

For our classification experiments, we have used a basic maximum entropy setup. The probability of a class label c given a segment of observations o is estimated as

$$P(c|o) = \frac{\exp(\sum_i \lambda_i f_i(c,o))}{\sum_{c'} \exp(\sum_i \lambda_i f_i(c',o))}$$

The features f_i, each of which measures some form of consistency between the audio in the segment and the label c are defined in Section 3. Training is done with Rprop [11] so as to maximize the regularized conditional data likelihood. Classification enumerates the possible segment labels and outputs the highest likelihood label.

2.2. Segmental CRF for Recognition

A graphical representation of a Segmental CRF if shown in Fig. 1. It is a "two-layer" model in which the observations in the bottom layer are linked to the label sequence in the top layer. Atomic observations at the frame level are grouped

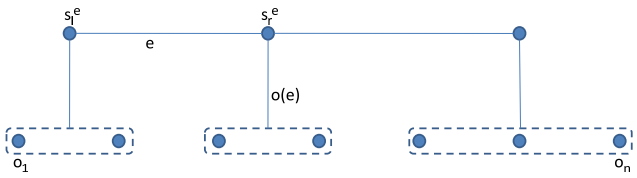

Fig. 1. A Segmental CRF.

together into segments with precisely defined boundaries, and the model is defined in terms of these segmentations.

Denote by $\mathbf{q}$ a segmentation of the observation sequences, for example in Fig. 1 where $|\mathbf{q}| = 3$. The segmentation induces a set of (horizontal) edges between the states, referred to below as $e \in \mathbf{q}$. One such edge is labeled e in Fig. 1 and connects the state to its left, s_l^e, to the state on its right, s_r^e. Further, for any given edge e, let $o(e)$ be the segment associated with the right-hand state s_r^e, as illustrated in Fig. 1. With this notation, we represent all features as $f_k(s_l^e, s_r^e, o(e))$. The conditional probability of a state sequence $\mathbf{s}$ given an observation sequence $\mathbf{o}$ for a SCRF is then given by

$$P(\mathbf{s}|\mathbf{o}) = \frac{\sum_{\mathbf{q}\ s.t.\ |\mathbf{q}|=|\mathbf{s}|} \exp(\sum_{e\in\mathbf{q},k} \lambda_k f_k(s_l^e, s_r^e, o(e)))}{\sum_{\mathbf{s}'} \sum_{\mathbf{q}\ s.t.\ |\mathbf{q}|=|\mathbf{s}'|} \exp(\sum_{e\in\mathbf{q},k} \lambda_k f_k(s_l'^e, s_r'^e, o(e)))}.$$

Training is done by gradient descent with a regularized conditional maximum likelihood objective function, and a description of the update equations and decoding recursions may be found in [1].

In [8], this model was adapted to large vocabulary continuous speech recognition by making its states correspond to words, and constraining the set of segmentations to those present in a lattice from a first-pass HMM recognizer. In this work, the states correspond to phones, and we consider all possible segmentations and labellings.

The features used in these experiments are based on a vector-quantized representation of the audio, using one or more VQ streams. The features used in the classification experiments are all binary - the presence or absence of a pattern in a VQ stream. Note that when segmentation is done, binary features may create a bias towards fewer segments since long spans of audio can be boiled down to one or two feature values of "1." Therefore, for the SCRF we use "counting" versions of the features - *how many times* a particular pattern is seen. This, combined with the inclusion of unigram features, causes each frame to be counted. A similar issue was dealt with in a generative context in [12] by using the probability ratio of phone to anti-phone models. The discriminative SCRF training procedure has a similar effect. The features are illustrated in Table 1 and described further below.

3. FEATURES USED

1. **Offset Feature:** The offset feature always has a value of "1" and there is one offset feature per class. Thus, the model can learn phone priors.

VQ Sequence: l m n o | a b c d e | q r s t	
Feature	**Values Extracted**
VQ	a,b,c,d,e,ab, bc, de
Prefix	a,ab,abc,abcd
Suffix	e, de, cde, bcde
Lead-in	o, no, mno
Follow-up	q, qr, qrs
Left-context	o, n, m, l, lm, mn, no
Right-Context	q, r, s, t, qr, rs, st
Left-boundary	oa
Right-boundary	eq

Table 1. Illustration of pattern-based features. The ngram level of the VQ and context features is 2. The VQ sequence of the hypothesized phone is shown between vertical bars.

2. **Length Features:** The length feature is of the form "the phone is X and the length is Y" for each phone and length combination up to a maximum length.
3. **VQ Ngrams:** Ngram count features are of the form "the phone is X and ngram pattern P is present." These are created for ngram patterns seen in the data from unigrams up to a maximum length.
4. **Left and Right Context Ngrams:** These features are the same as the previous, except that the ngram counts are extracted from a window of 4 frames immediately to the left and right of the segment boundaries.
5. **Prefix, Suffix Patterns:** These features are of the form "the phone is X and the ngram Y is a prefix/suffix of the segment." We used prefixes and suffixes up to length 4.
6. **Lead-In, Follow-up Patterns:** Lead-in and follow-up features are similar to the prefix and suffix features, but refer to the audio immediately to the left or right of a segment. They are of the form "the phone is X and the ngram Y terminates/begins immediately to the left/right of the segment." We used patterns of up to length 3.
7. **Cross-Boundary Patterns:** These features are of the form "the phone is X and the pair of VQ symbols straddling the the segment boundary is YZ."
8. **Language Model Features:** The use of the left and right context features gives clues about the identity of the surrounding phones, and thus implicitly models phone bigrams. To understand this effect better, we introduced explicit language modeling features - of the form "the phone to the left/right is X and the hypothesized phone is Y." By using the phone values from the transcription, we can estimate the amount of potential gain with acoustic context features.

4. EXPERIMENTAL RESULTS

4.1. TIMIT Data and Tasks

We present results for the TIMIT [13] phone classification and recognition tasks. We use the defined development and core

Quantization Level	1gram	2gram	3gram
128	37.6%	34.8%	34.8%
512	33.1	31.7	31.8
2048	29.7	29.6	29.5
4096	29.0	29.0	29.0
8192	28.5	28.8	28.9

Table 2. Effect of VQ granularity and level of ngram features on classification error rate.

test sets. We follow the conventional [14] use of 48 phone classes for modeling, and map these to the standard 39 for scoring. Acoustic processing was done in 25 ms frames extracted every 10 ms. We used MFCC coefficients, deltas and double-deltas based on a 40-channel Mel filterbank. Prior to vector quantization, CMS was applied. Vector quantization was done via k-means clustering with diagonal covariance gaussians. Unless otherwise specified, the MFCCs, deltas and double-deltas were quantized together. Also unless otherwise specified, results are reported on the dev set, with results on the core test set being explicitly presented at the end.

4.2. Classification Results

Table 2 shows the effects of different levels of VQ granularity and VQ ngram order. Note that the ngram order is inclusive, so e.g. 2-gram features include 1-grams. Three types of features are used, which we term the "base features:" offset, length and VQ ngrams. We see that two effects are visible: performance continually improves as the granularity is made finer; and, whereas with low-granularity VQ streams high-order ngrams are important, with a large number of codewords, they are not necessary.

Table 3 shows the effectiveness of adding *individual* features to the base feature set. We see significant improvement from all the features. In particular, the use of the left and right context reduces the error rate by more than 4% absolute.

Given the large effect of left/right context features, we investigated their power relative to oracle language modeling features. Note that the acoustic context features are at *segment boundaries*, and we can therefore expect them to be a reasonable characterization of the surrounding phones. This is in contrast to the use of surrounding frames in a frame-wise system. Table 4 shows this for the 8192 level VQ system of the previous table. Using all features except those which examine the acoustics surrounding the segment boundaries results in a 27.8% error rate. Adding the oracle LM features reduces this by 7.3% absolute. Using instead the surrounding context features reduces the error rate by 4.2% - over half the oracle gain if we were given correct identities of the surrounding phones. Thus the acoustics immediately surrounding the boundaries may be an inexpensive surrogate for LM features (the use of bigram features increases the computational complexity by a factor proportional to the number of phones).

In Table 5, we present the effect of combining multiple

Feature	2048	4096	8192
Base	29.6%	29.0%	28.5%
+ Cross-Boundary	29.2	28.6	28.5
+ Prefix-Suffix	28.1	28.2	27.8
+ Lead-Follow	26.9	26.9	26.6
+ Left/Right Context	24.5	24.5	24.2
All	24.2	24.0	23.6

Table 3. Effect of adding individual features on classification error rate. 2048 and 4096 level VQ systems use 2gm features; 8192 level VQ system uses 1gm features.

Features	Dev PER
All But AM Context	27.8%
+LM Context Only	20.5
+AM Context Only	23.6
+ AM and LM Context	18.3

Table 4. Acoustic context features compared with language modeling features.

streams of information, and results on the core test set. We achieve results competitive with some of the best discriminatively trained classification systems using continuous features (Section 4.4).

4.3. Recognition

Having explored the use of our discrete segment features in classification, we turn now to their use in recognition. As noted in [12] segmental models can introduce bias, e.g. towards results with fewer segments. In this work we rely on the use of length features, unigram ngram features (whose count is constant across frames), and the discriminative nature of SCRF training to provide reasonable segment lengths. In decoding, we have also found it beneficial to use an extra "insertion penalty" as in conventional systems.

On the assumption that energy change might help in signaling phone boundaries, we have quantized c0 and delta-c0 into a separate stream. Table 6 shows initial recognition results using a 4096 level VQ stream with the base features, all features, and an additional c0 stream.

4.4. Discussion

In the classification task, the lowest error rate we have achieved is 21.7%, which compares well with, e.g. 24.6% for MMI-HMMs [15], 21.7% for HCRFs [15], 20.8% for a continuous HCRF with distribution constraints [16], 21.1% for large-margin GMMs [17], 21.0% for multiple segment models [9], and 23.0% for segment NNs [18]. We note that committee classifiers have achieved 16.8% [19].

The best recognition result obtained in these experiments is 33.1%. This is in line with results for discrete monophone

Streams Used	Dev	Core Test
8192	23.6%	25.3%
8192 + 4096	22.7	23.7
8192 + 4096 + 2048	21.6	22.1
8192 + 4096 + 2048 + c0	21.2	21.7

Table 5. Effect of adding multi-granularity VQ streams.

Features and streams	Dev PER	Core PER
4096	33.2%	34.1%
4096 + c0	33.0	34.0
4096 + All features	32.4	33.1

Table 6. Recognition results.

models from the literature, e.g. 35.1% from [14], and some other systems, e.g. 35.9% for a monophone segment model [20]; 30.5% for diphone segment model [20]; 32.1% for a frame-level CRF with articulatory features [21], or 30.1% for large margin HMMs [17]. However, there is a significant gap from the best current results, e.g. [22] which achieves 19.7% using deep neural networks and discriminative input features. We believe that the addition of continuous features and context dependent acoustic units will help close this gap.

5. CONCLUSION

This paper has described some first steps towards using direct modeling to create segment classifiers and phone-level detectors. We define a set of discrete segmental features suitable for use in both maximum entropy segment classifiers, and in segmental CRF phone detection. We find that the acoustics in the region surrounding the segment boundaries can be effectively exploited using these features, and produces a large fraction of the gain that could be obtained if we actually knew the surrounding phone identities. The maximum entropy classification setup produces excellent phone classification results, and we have demonstrated the feasibility of bottom-up SCRF phone recognition.

Acknowledgments

We thank Mark Gales for many insightful comments on segmental models and associated discriminative training methods.

6. REFERENCES

[1] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*, 2009.

[2] M. I. Layton and M. J. F. Gales, "Augmented statistical models for speech recognition," in *Proc. ICASSP*, 2006.

[3] S-X. Zhang, A. Ragni, and M. J. F. Gales, "Structured log linear models for noise robust speech recognition," in *IEEE Signal Processing Letters*, To Appear.

[4] J. Kao, G. Zweig, and P. Nguyen, "Discriminative duration modeling for speech recognition with segmental conditional random fields," in *ICASSP*, 2011.

[5] G. Heigold, G. Zweig, X. Li, and P. Nguyen, "A flat direct model for speech recognition," in *Proc. ICASSP*, 2009.

[6] K. Demuynck, D. Seppi, D. Van Compernolle, P. Nguyen, and G. Zweig, "Integrating meta-information into exemplar-based speech recognition with segmental conditional random fields," in *ICASSP*, 2011.

[7] D. Seppi, K. Demuynck, and D. Van Compernolle, "Template-based automatic speech recognition meets prosody," in *Interspeech*, 2011.

[8] G. Zweig, P. Nguyen, D. Van Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, S. G.S.V.S., S. Bowman, and J. Kao, "Speech recognition with segmental conditional random fields: A summary of the JHU 2010 summer workshop," in *ICASSP*, 2011.

[9] Andrew K. Halberstadt and James R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *EuroSpeech*, 1997.

[10] Hua Yu and Alex Waibel, "Integrating thumbnail features for speech recognition using conditional exponential models," in *ICASSP*, 2004.

[11] M. Reidmiller, "Rprop - description and implementation details," Tech. Rep., University of Karlsruhe, January 1994.

[12] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, 2003.

[13] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *DARPA Speech Recognition Workshop*, 1986.

[14] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, 1989.

[15] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," in *Interspeech*, 2005.

[16] D. Yu, L. Deng, and A. Acero, "Hidden conditional random field with distribution constraints for phone classification," in *Interspeech*, 2009.

[17] F. Sha and L. Saul, "Large margin gaussian mixture modeling for phonetic classification and recognition," in *ICASSP*, 2006.

[18] S. Zahorian, P. Silsbee, and X. Wang, "Phone classification with segmental features and binary-pair partitioned neural network classifier," in *ICASSP*, 1997.

[19] H. Chang and J.R. Glass, "Hierarchical large-margin gaussian mixture models for phonetic classification," in *ASRU*, 2007.

[20] J.R. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-fased speech recognition," in *ICSLP*, 1996.

[21] J. Morris and E. Fosler-Lussier, "Discriminative Phonetic Recognition with Conditional Random Fields," in *HLT-NAACL*, 2006.

[22] A. Mohamed, T. Sainath, G. Dahl, B. Ramabhadran, G. Hinton, and M. Picheny, "Deep belief networks using discriminative features for phone recognition," in *ICASSP*, 2011.