Transition features for CRF-based speech recognition and boundary detection

Spiros Dimopoulos [1], Eric Fosler-Lussier [2], Chin-Hui Lee [3], Alexandros Potamianos [1]

[1] *Dept. of Electronics and Computer Engineering, Technical University of Crete, Chania 73100, Greece*

[2] *Dept. of Computer Science and Engineering, 2015 Neil Avenue, Columbus, OH 43210-1277, USA*

[3] *School of Electrical and Computer Engineering, 777 Atlantic Drive NW, Atlanta, GA 30332-0250, USA*

{sdim,potam}@telecom.tuc.gr fosler@cse.osu.edu chl@ece.gatech.edu

***Abstract*—In this paper, we investigate a variety of spectral and time domain features for explicitly modeling phonetic transitions in speech recognition. Specifically, spectral and energy distance metrics, as well as, time derivatives of phonological descriptors and MFCCs are employed. The features are integrated in an extended Conditional Random Fields statistical modeling framework that supports general-purpose transition models. For evaluation purposes, we measure both phonetic recognition task accuracy and precision/recall of boundary detection. Results show that when transition features are used, in a CRF-based recognition framework, recognition performance improves significantly due to the reduction of phone deletions. The boundary detection performance also improves mainly for transitions among silence, stop, and fricative phonetic classes.**

I. Introduction

Hidden Markov Model (HMM)-based speech recognition systems use state models. Audio features are used to compute the likelihood of a state for each speech model, but no exclusive information about phonetic transitions is extracted from the audio signal [1]. A simple statistical (Markovian) model of the transition probabilities between states and models is used. In some cases, the state transition probabilities are ignored by setting them all equal. Our goal is to explicitly model the transitions between speech units and states with features that are extracted directly from the speech signal.

The benefit of including transition information into the state models has been recognized in past work. In [2], an acoustic feature set that captures the dynamics of the speech signal at the phoneme boundaries was introduced in combination with the traditional acoustic feature set representing the periods of speech that are assumed to be quasi-stationary. In [3], a extended hidden Markov model that integrates generalized dynamic feature parameters into the model structure is developed and evaluated. Incorporating transition information is also critical for segmental-based recognition techniques [4], [5].

In this paper, we use the Conditional Random Fields (CRF) speech recognition framework that supports the coupling of transitions between phonetic units and states with features [6]. We investigate various spectral- and energy-based distance metrics, as well as, time derivatives of phonological descriptors and MFCCs as transition features. The extended CRF models are trained and tested for a phone recognition and boundary detection task.

II. Transition features

In order to improve speech recognition performance, one can model explicitly transitions between adjacent speech units. Such models can be trained using features from both spectral and time domains. The time derivatives (deltas) of features associated with particular states can also be used as indicators of a change in state (and thus are transition features). We examine these two classes of features below.

A. Spectral and energy domain

A common feature for this task is a measure of spectral change. We have already used this feature to adapt our system to the rate of change of speech signal in [7]. In [8], a similar feature set was used for automatic segmentation in a speech synthesis application. We use the Mel-Scale Spectral Magnitude to compute the spectral region differences of 20 filterbank channels. We then combine these sub-spectral differences with the product rule and equal weights. The spectral change metric for frame i is computed by the equation (average over three frames):

$$D(i) = \frac{\sum_{j=i-1}^{i+1} \prod^{K} d_j(k)}{3} \quad (1)$$

where $d_j(k)$ the distance for j-th frame and k-th spectral region, and $K = 20$ is the number of spectral regions used.

A similar feature is the Spectral Flux (Fss) [9]. The Spectral Flux is the difference between the amplitudes of successive magnitude spectra:

$$F_{ss}^{0}(i) = \sum_{k=0}^{K} [M_i(k) - M_{i-1}(k)]^2. \quad (2)$$

where $M_i(k)$ and $M_{i-1}(k)$ are the magnitudes of the spectra for frames i and $i-1$. Fss measures the amount of spectral change between successive frames. We derived a smoothed flux measure by averaging over neighboring flux measurements:-

$$F_{ss}(i) = \frac{\sum_{j=i-1}^{i+1} F_{ss}^{0}(j)}{3} \quad (3)$$

Another important group of spectral features is the Spectral Centroid difference, the Spectral Roll-off difference and the Zero-Crossing Rate difference [9]. The Spectral Centroid (Css)

is the frame-to-frame difference of the center of mass of power spectrum

$$C_{ss}(i) = \frac{\sum^{K} kP_i(k)}{\sum^{K} P_i(k)} \quad (4)$$

where $P_i(k)$ is the power spectrum for frame i and frequency k, and K is the total number of frequency bins. The Spectral Roll-off (Rss) is the frequency below which the 95% of the power spectrum is concentrated. Finally, the Zero-Crossing Rate (Zss) is the rate of sign changes (positive to negative and back) of a signal and can be computed in the time-domain. The (smoothed) time difference of each of these features was used as follows:

$$X_{ssd}(i) = \frac{\sum_{j=\{1,2\}} X_{ss}(i+j) - X_{ss}(i-j)}{2} \quad (5)$$

where $X_{ssd}()$ is the difference feature and $X_{ss}()$ one of the previously described spectral features for frame i.

Finally, the frame-to-frame Energy difference of the signal was also used for boundary detection. The same regression formula was used as in (5).

To evaluate the proposed features, a linear classifier which worked as a detector/rejector of transition regions was used for each feature. Adjusting the operation point, for the boundary detector the optimal hits to false positive ratio is reported, while for the boundary rejector the optimal true negative to miss ratio is reported in Table I. Overall, spectral change and spectral flux perform better as rejectors of frames as possible transitions, especially for frames in the same phonetic class, e.g., STOP→STOP. Css, Rss, Zss and energy difference are better detectors. Best results were obtained for transitions between VOW→{s,z}, {s,z}→VOW and n→{s,z}

Features	Detector Ratio	Rejector Ratio
Spectral change	2.56	**5.78**
Spectral flux	1.31	**14.45**
Spectral centroid diff	**6.72**	2.35
Spectral roll-off diff	**4.21**	2.94
Zero crossing rate diff	**10.16**	2.03
Energy diff	**3.53**	1.24

TABLE I
SPECTRAL AND ENERGY FEATURE EVALUATION

B. Phonological and MFCC deltas

Phonological features have long been used to describe whether phonological attributes of segments, such as the consonant manner, place of articulation and voicing, sonority, or vocalic attributes, are present within a speech frame ([6], [10] inter alia). These attributes are associated with a group of phonetic units; each unit can be thought of as a bundle of features. The relationship between segment boundaries and phonological features can be complex: while some features can extend across boundaries (as in the nasalization of vowels), many features will transition in unison at segment boundaries. The degree to which features transition in concert depends particularly on the type of segmental transition.

Deltas of phonological features can be used to estimate the rate of change of the phonological attributes: High values of phonological deltas indicate a phonological attribute transition. We used phonological deltas as transition features, accepting the risk that we will have a small increase in false positive detection of boundaries (and consequently in the insertions of the final recognition task).

The common MFCC features vector was computed with a frame-rate of 10 msec and a window size of 25 msec. MFCC deltas is a commonly used group of features that estimate spectral change in time, in the transformed cepstral domain. MFCC deltas were also used as transition features.

III. CRF RECOGNITION USING TRANSITION FEATURES

The Conditional Random Field (CRF) framework is a successful integration tool for combining features that are highly correlated and of different quality [7], [11]. Recent advances in the framework allow the inclusion of phonetic transition boundary clues as transition features. We use the CRF recognition framework to test our boundary detection methods.

A. State and transition functions in CRF

CRF models are exponential models that use functions of the input features and a trainable weighting scheme of these functions to model the phonetic units. State functions are associated with states and state features. These are used to compute the likelihood of being in a certain state. In addition, transition functions are associated with transitions between states and transition features. The posterior probability $P(y|x)$ of a phonetic unit label sequence y given an input feature sequence x is given by:

$$P(y|x) \propto exp \sum_{i} (S(x,y,i) + T(x,y,i)) \quad (6)$$

where

$$S(x,y,i) = \sum_{j} \lambda_j s_j(y,x,i) \quad (7)$$

and

$$T(x,y,i) = \sum_{k} \mu_k t_k(y_{i-1}, y_i, x, i) \quad (8)$$

Each state feature function $s(y,x,i)$ is associated with a phonetic unit label and an input state feature and also has an index pointing to a position in the feature sequence. Similarly, each transition feature function $t(y_{i-1}, y_i, x, i)$ is associated with a phonetic unit transition and a transition input feature and also has an index in the feature sequence. Trainable weights λ and μ learn the importance of the association of each phonetic unit label or transition with the state or feature function in the final probability calculation.

B. Using transition features to improve the boundary detection

Prior work in CRF phonetic recognition has either ignored or used a simplified approach in the transition function implementation. In [6], the transition functions were binary, evaluating to 1 when the phonetic unit label pair matched the values for the defined function and 0 otherwise. This left out any transition clues that where present in the input and let the Viterbi decoding decide which transitions maximized the final probability of the sequence. In [11], a feature set was

Setup	State features	Transition features
Baseline	48 Phonological features	No transition features
DPhn	48 Phonological features	48 Phonological Delta (1st order)
BndF	48 Phonological features	6 boundary features (see Table I)
DPhn + BndF	48 Phonological features	48 Phonological Delta (1st order) + 6 boundary features
DPhn + BndF + DMFCC	48 Phonological features	48 Phonological Delta (1st order) + 13 MFCC delta (1st order) + 6 boundary features

TABLE II
DESCRIPTION OF BASELINE AND DIFFERENT EXPERIMENTAL SETUPS

	CV Set		Core Test Set					Ext Test Set
	Acc %	Train Iters	Acc %	Corr	Del	Subs	Ins	Acc %
Baseline	71.4	25	69.11	4597	941	976	95	70.25
DPhn	72.46	11	70.42	4705	772	1037	118	71.46
BndF	72.02	19	69.53	4652	849	1013	123	70.84
DPhn + BndF	73.02	2	70.77	4773	692	1049	163	71.76
DPhn + BndF + DMFCC	**73.72**	3	**71.51**	4825	647	1042	167	**72.32**

TABLE III
RECOGNITION PERFORMANCE FOR DIFFERENT EXPERIMENTAL FEATURE SETUPS

used for both state and transition features. This feature set was not optimized for boundary detection, thus allowing only a small increase in overall recognition performance. Explicit boundary detection (using a single MLP detector to detect segmental boundaries) was found to be mildly effective, but an expensive transition feature for CRF transitions in [12]. In this work, we examine a wide range of boundary features which are designed to detect as many boundaries as possible without adding a considerable amount of insertions to the system.

IV. EXPERIMENTAL SETUP

We used as baseline the CRF recognition system using 48 phonological features as state features and no transition features as in [11]. We used 4 different setups to compare the transition features and their combinations. The experimental setups are described in Table II.

Phonological features were computed using Multi-Layer Perceptron (MLP) detectors of phonological attributes and trained using 13-D PLP coefficients plus velocity and acceleration. MLPs used 2000 hidden units.

Boundary features were computed as described in section II. The frame-rate of analysis was 10 msec and the window size 25 msec. The same window parameters were used for MFCC calculation.

The delta features where computed with a 7-frame window size and with the following regression formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta}(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{9}$$

with $\Theta = 3$ and c_t the static features [13].

V. PERFORMANCE REPORTING AND ANALYSIS

We report two different performance indicators. The first is the final recognition performance of the setup. This is an indicator of how well our experimental boundary feature setup has done in recognizing phonetic units. The second is how well our setup has done in detecting the transition boundaries of phonetic units.

A. Recognition results

Our first set of performance indicators are the overall recognition results of different setups. The recognition is performed for 48 phonetic units which are reduced to 39 during performance evaluation. We report for 3 sets: Cross Validation (CV), Core Test and Extended Test as in [11]. Results are shown in Table III. By using the phonological deltas (DPhn) we got a marginally significant improvement in accuracy. In contrast, when we used the six boundary features alone (BndF), the improvement was not significant. Then when we used both phonological deltas and boundary features (DPhn + BndF) we got a better accuracy from the previous two experiments, as expected. Finally when we used all available transition features - phonological deltas, boundary features and MFCC deltas (DPhn + BndF + DMFCC) - we got the best accuracy. Note that the improvement is due to the significant reduction in deletions (by over 20%). Also by adding more transition features, the training process converges with only a couple of iterations.

B. Boundary detection results

In addition to recognition performance, we can see how well the CRFs directly detect segment boundaries. We report the overall boundary detection performance, i.e., the detection ratio for transitions between two phonetic units in terms of precision and recall for the extended test set. These results in Table IV offer an overview of the detector performance. Two tolerance levels in the detection of transition boundaries are reported: 10 msec (strict) and 20 msec (normal). One can see that when using phonological deltas, a slight increase in recall is achieved with a matching decrease in precision. When using boundary features, we get an increase in recall without any loss in precision. When using both phonological

	NAS ↔ STOP		VOW ↔ FRIC		VOW↔ STOP		FRIC ↔ SIL		STOP ↔ SIL	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Baseline	0.70/0.69	0.64/0.86	0.92/0.89	0.94/0.93	0.82/0.92	0.95/0.97	0.79/0.75	0.80/0.75	0.66/0.73	0.52/0.49
DPhn	0.71/0.74	0.65/0.87	0.92/0.89	0.93/0.93	0.83/0.91	0.95/0.96	0.78/0.75	0.80/0.77	0.64/0.70	0.56/0.60
BndF	0.71/0.76	0.65/0.86	0.92/0.89	0.94/0.94	0.83/0.91	0.96/0.96	0.80/0.73	0.81/0.78	0.69/0.76	0.57/0.65
DPhn + BndF	0.72/0.71	0.70/0.96	0.92/0.87	0.94/0.95	0.84/0.91	0.96/0.96	0.79/0.72	0.81/0.78	0.65/0.72	0.58/0.68
DPhn + BndF + DMFCC	0.74/0.75	0.71/0.88	0.92/0.89	0.94/0.95	0.84/0.91	0.96/0.97	0.80/0.72	0.83/0.78	0.68/0.73	0.59/0.70

TABLE V
EXAMPLES OF BROAD PHONETIC CLASS BOUNDARY DETECTION PERFORMANCE

deltas and boundary features we get a complementary effect, recall increases significantly with a small decrease in precision. Finally the addition of MFCC deltas provides a negligible gain in recall. The detailed performance for transitions between broad phonetic classes (BPC) are reported in Table V for the extended test set. The phonetic units are grouped into 5 classes, namely: vowels and semi-vowels (VOW), fricatives (FRIC), nasals-flaps (NAS), stops (STOP), silence (SIL). Detection results (precision/recall) for each experimental setup and transitions between these BPC are reported for the strict 10 msec window. The first value in each cell of Table V is the precision/recall ratio of the transition boundary of the left phonetic class to the right as presented in the table (while the second value is for the right to left transition). Overall, by adding transition features into the CRF framework, boundary detection improves significantly especially among the SIL, STOP and FRIC phonetic classes. It seems that these phonetic classes transitions get the highest complementary effect from the different groups of transition features, so they finally increase their recall without losing precision.

Tolerance:	10 msec		20 msec	
	Precision	Recall	Precision	Recall
Baseline	0.89	0.78	0.955	0.855
DPhn	0.875	0.795	0.95	0.88
BndF	0.89	0.795	0.955	0.87
DPhn + BndF	0.88	0.81	0.945	0.89
DPhn + BndF + DMFCC	0.88	**0.815**	0.945	**0.895**

TABLE VI
OVERALL BOUNDARY DETECTION PERFORMANCE

VI. CONCLUSION

In this paper, we proposed the use of features extracted from the speech signal to detect the boundaries of transitions between adjacent phonetic units. In addition, the CRF framework was extended to incorporate such transition features. Overall, we showed that spectral distance metrics can help reject erroneous transition hypothesis, while energy-based metrics are good detectors of transitions. By incorporating transition features into CRF-based speech recognition a moderate, yet significant, improvement in phone accuracy was achieved due to the reduction of deletions. For the CRF-based phonetic boundary detection task, recall increased significantly when transition features were used, especially, for transitions among the silence, stop, and fricative phonetic classes.

This is a first step towards integrating transition features into CRF-based speech recognition. A variety of features extracted from the audio signal can be potentially used for boundary detection. In addition, one may explicitly create and train transition models of broad phonetic classes and use their likelihood scores as transition features, in the future.

ACKNOWLEDGMENTS

This work was supported in part by NSF ITR grant IIS-0427413. In addition, Spiros Dimopoulos was supported by an Onassis Foundation Scholarship and by a Technical University of Crete Research Travel Grant.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[2] M. Omar, M. Hasegawa-Johnson, and S. Levinson, "Gaussian mixture models of phonetic boundaries for speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2001, pp. 33–36.

[3] C. Rathinavelu and L. Deng, "Use of generalized dynamic feature parameters for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 232–242, May 1997.

[4] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From hmms to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 360–378, 1995.

[5] J. Chang, "Near-miss modeling: A segment-based approach to speech recognition," Ph.D. dissertation, MIT, 1998.

[6] J. Morris and E. Fosler-Lussier, "Further experiments with detector-based conditional random fields in phonetic recognition," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, April 2007, pp. IV–441–IV–444.

[7] S. Dimopoulos, A. Potamianos, E.-F. Lussier, and C.-H. Lee, "Multiple time resolution analysis of speech signal using mce training with application to speech recognition," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 3801–3804.

[8] Y.-J. Kim and A. Conkie, "Automatic segmentation combining an hmm-based approach and spectral boundary correction," in *Proc. 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[9] M. Lee, "Feature extraction toolbox: A matlab tool set for speech analysis," Georgia Institute of Technology, 2008.

[10] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[11] E. Fosler and J. Morris, "Crandem systems: Conditional random field acoustic models for hidden markov models," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, March-April 2008, pp. 4049–4052.

[12] Y. Wang, "Integrating phone boundary and phonetic boundary information into ASR sytems," Master's thesis, Dept. of Computer Science and Engineering, The Ohio State University, 2007.

[13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Microsoft Corporation, 1995-1999.