FOCUSED STATE TRANSITION INFORMATION IN ASR

Chris Bartels and Jeff Bilmes

Department of Electrical Engineering
University of Washington, Seattle
{bartels,bilmes}@ee.washington.edu

ABSTRACT

We present speech recognition graphical models that use "focused evidence" to directly influence word and state transition probabilities in an explicit graphical-model representation of a speech recognition system. Standard delta and double delta features are used to detect loci of rapid change in the speech stream, and this information is applied directly to transition variables in a graphical model. Five different models are evaluated, and results are given on the highly mismatched training/testing condition tasks in Aurora 3.0. The best of these models gives an average 8% reduction in word error rate over baseline, significant at the 0.05 level.

1. INTRODUCTION

Conventional hidden Markov model (HMM) based automatic speech recognition (ASR) systems are composed of a chain of pairs of random variables, where each pair comprises a hidden "state" variable and its associated observation variable. These hidden variables often use a single integer value to simultaneously represent a variety of information — this includes position within a word or sentence, word identity, lexical variant, word history, and so on. The resulting state transition table is thus not only a set of conditional probabilities, but it is also a representation of the allowed sequences of these complex states. Often, the hidden information is hierarchically structured (forming essentially a hierarchical HMM) where word, sub-word, state, and substate are represented separately but are flattened into a single network before recognition takes place.

An explicit graphical model (GM) representation of a speech recognition system, on the other hand, expresses this same information as a diverse network of latent random variables. Each of these variables has a straightforward meaning and a simple relationship to the other variables in the graph ,and many of these relationships are deterministic. For example, in Figure 2(a) there are separate variables modeling the word, word transition, position within the word, state transition, state, and acoustic observation [1, 2]. Such a representation exposes high-level information that is normally flattened into a single hidden variable and transition matrix. As such, this gives us the opportunity to "focus" highly tuned transformations of the speech signal directly on high-level portions of the speech recognition system, rather than indirectly via the lowest-level (or a flattened) state variable using either an appendage to or substitution in a feature vector. We have called this the *focused approach*, and have successfully applied this idea in [3], where acoustics are used to directly influence the word vs. silence hypothesis in an ASR system.

In this work, we introduce a new ASR model under the focused approach where acoustic/spectral transition information is used to directly influence hidden variables in a GM-based ASR system that indicate various forms of transition, namely inter-word transition and intra-word (or inter word-constituent) transition. Specifically, we focus standard delta and double-delta features directly on transition variables in addition to using it as an appendage in a regular MFCC-based feature vector. We apply this approach to the Aurora 3.0 noisy-speech corpus, in the highly-mismatched training/testing conditions, and find that we can achieve significant word-error (WER) reductions relative to a baseline state-of-the-art system.

Clearly, the use of delta and double-delta information in ASR is not new — what is new here, rather, is the manner in which it is employed. Indeed, the use of transition information has a long history of improving automatic speech recognition accuracy. In [4] polynomial expansion coefficients were used as part of a speaker verification system and [5] used delta features (calculated from a simple difference) to weight distances in a dynamic time warping isolated-word recognizer. The work in [6] used delta features as an augmentation of the feature vector in an HMM recognizer which is the manner that they are predominantly used today. It was demonstrated in [7] that delta features appended to the feature vector help in noisy conditions and in particular under the Lombard effect. Perceptual experiments have shown that transitional periods in speech play a role in human speech perception that may be more significant than stationary periods [8]. Double-delta features have been used since [9, 10]. Moreover, work such as [11] and [12] place the statistical focus of a speech recognizer directly on these transitional regions. Without a doubt, the use of time-derivative features is now a necessary component in

This work was supported by ONR MURI grant N000140510388 and by NSF grant IIS-0093430.

any modern speech recognition system.

The rest of this paper presents our new models that have the potential to take even better advantage of this information: Section 2 describes our general approach, Section 3 overviews our Aurora 3.0 setup, Section 4 describes each of our new graphical models in detail, Section 5 give results, and, lastly, Section 6 concludes.

2. FOCUSED EVIDENCE TRANSITION MODELS

Hidden variables that represent transition in an explicit GM-based ASR system are bound to indicate either acoustic signal change or at the very least indicate a forced evolution of the model towards the completion of an utterance. Consider, for example, the two binary indicator variables *word transition* $\mathcal{W}_t^{tr}$ and *state transition* $\mathcal{S}^{tr}$ in Figure 2(a) — the variable $\mathcal{W}_t^{tr}$ (resp. $\mathcal{S}^{tr}$) indicates movement from one word (resp. sub-word state) to the next. Normally, the influence that the acoustics has on these transition variables must occur indirectly via the state variable. This means that for a transition event, from say state i to j, to be encouraged, the acoustic feature vectors over one length-ℓ time region ($\mathcal{O}_\tau^s$, $\tau = t-\ell, \ldots, t-1$) should be correlated with one state value (say $\mathcal{S}_\tau = i$), and the vectors over the next length-r region ($\mathcal{O}_\tau^s$, $\tau = t, \ldots, t+r-1$) should be correlated with another state value ($\mathcal{S}_\tau = j$). This approach, which is also the case in standard HMM-based ASR systems, need not be the most efficient way to transfer information from acoustic transitions to the transition events with which they should ideally correlate.

A more focused (and likely more efficient) approach is to have acoustic transition information directly influence the transition events in a speech recognition system, something that might also improve the alignments represented by the Viterbi decodings. This idea can be easily done in the GM-framework as shown in Figures 2(b) through Figures 2(f). Of course, there are many possible signal-processing choices for a measure of acoustic transition information to be used as additional observations. In this work, we choose first to evaluate standard delta and double-delta features in this manner, already used in an ASR system via the state variable. In other words, we use delta and double-delta features both to augment the standard MFCC-based feature vector, and also to directly influence transition events, and we do so for the following reason: Figure 1 demonstrates the behavior of the delta features over an instance of the German word "sieben". A line showing the sum of the vector of the magnitude of first order deltas generated from 13 MFCC coefficients is superimposed over a spectrogram of the audio waveform. One can observe peaks in the delta features at spectral changes, phonetic boundaries, and (at least on Aurora 3.0) word boundaries. Therefore, when wishing to directly influence either word or state transition in an ASR model, delta and double delta features (and specifically peak detection) are likely to be beneficial. Note that we expect double deltas to be useful because a small value for the second derivate indicates a peak in the first derivative.

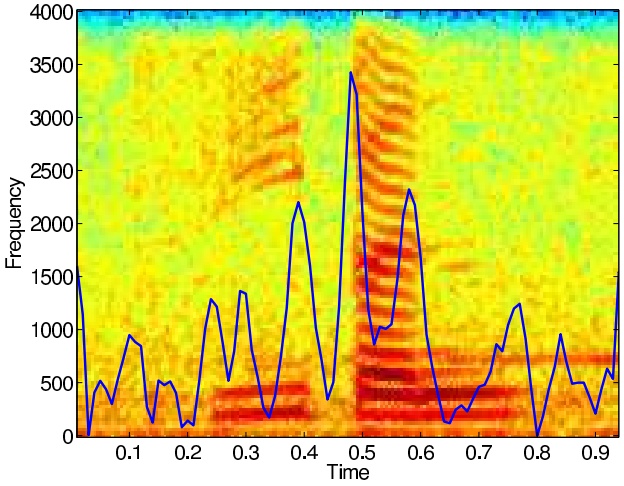
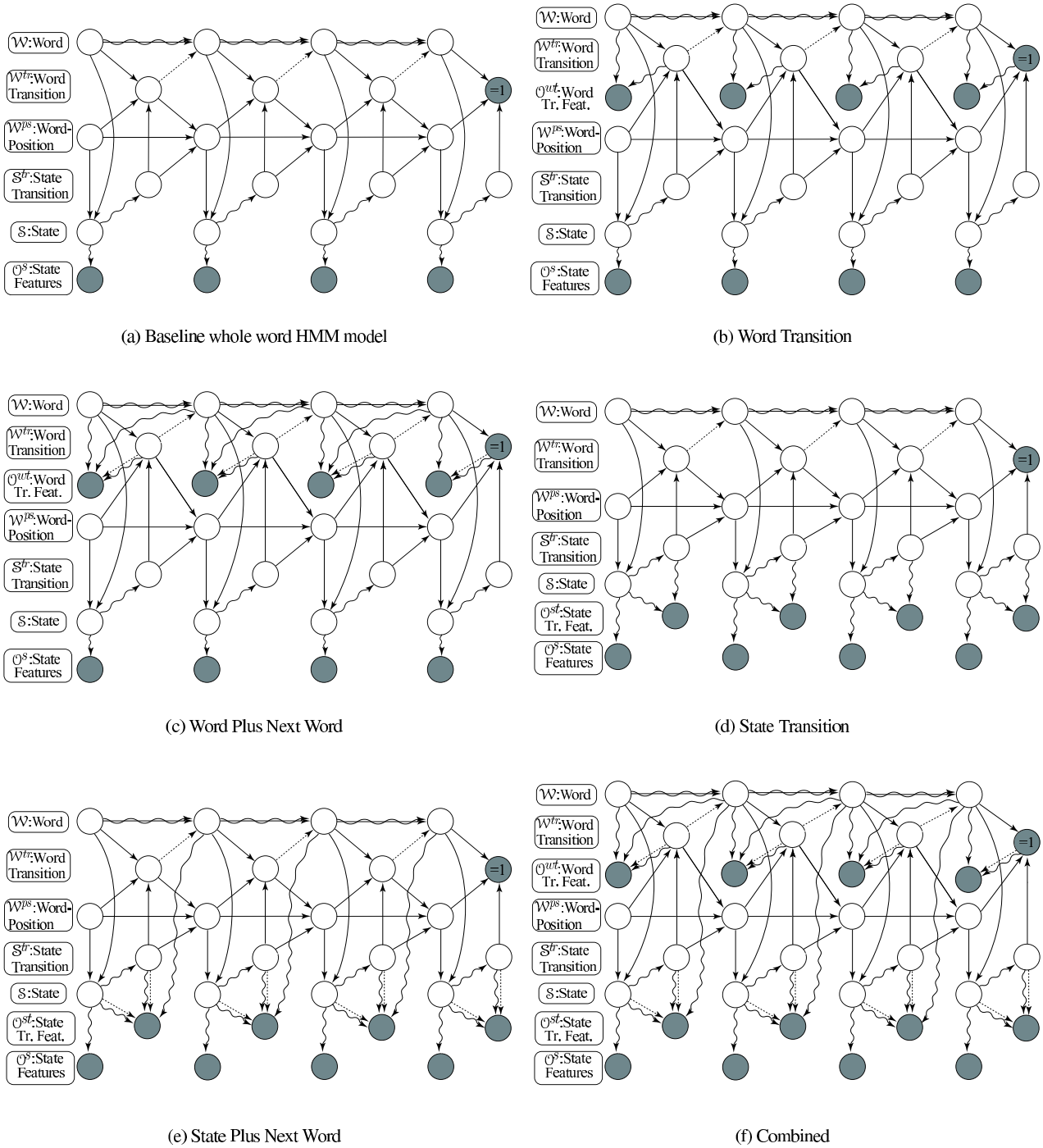


Fig. 1. Sum of delta magnitudes overlaid on the spectrogram of the German word "sieben".

One possible criticism of these models is that they incorporate delta features at multiple observations, and thus creates an unnormalized product model. The use of such a model could loose some of the sufficient conditions that are theoretically available during parameter training which guarantee convergence to a local maxima of the likelihood function. We have empirically found, however, that likelihood values continue to increase monotonically when training these models using standard expectation-maximization (EM) training. Interestingly, this issue is not dissimilar to the state of affairs in standard HMM-based speech recognition training, where successive features vectors are constructed from windows of the underlying speech signal that overlap by 15 out of the typically 25ms window width. Moreover, the use of deltas in a feature vector to begin with doubly presents the acoustic information to the HMM system, since the delta features are a deterministic function of the original features. Arguably, in such systems acoustic evidence is already "double counted" but we continue to see monotonic likelihood increases. Lastly, training using a likelihood cost criterion is not ideal either, as we really desire a discriminatively formed model — a wrong model from a generative perspective might work quite well when used as a classifier [2]. In any event, we use these models as is, and agree that more theoretical work is needed in this area to justify these empirical successes.

3. CORPUS AND EXPERIMENTAL SETUP

We use the Aurora 3.0 corpus for all experiments in this paper. This corpus has digit recognition tasks in Danish, Finnish, German, and Spanish recorded under varying noise conditions. Danish and German have 11 words, while Finnish and Spanish have 10. Aurora 3.0 has three types of training/testing conditions: well-matched, medium-matched, and



(a) Baseline whole word HMM model

(b) Word Transition

(c) Word Plus Next Word

(d) State Transition

(e) State Plus Next Word

(f) Combined

Fig. 2. Dynamic Bayesian Networks that use "focused" evidence to predict state transitions. Solid edges represent deterministic relationships, wavy edges are probabilistic relationships, and dashed edges are switching parents [13] whose values select a subset of the other edges. Hollow circles are hidden variables and filled circles are observed.

highly-mismatched. We choose to evaluate the quality of our systems using the latter case. The reason for this is because highly mismatched train and test conditions are generally perceived as the most realistic environment an automatic speech recognition (ASR) system must operate in.

The features are 13 dimensional MFCCs created at 10ms intervals using a 25ms Hamming window and a bank of mel-filters between 64 Hz and 4000 Hz. 13 delta features and 13 double delta features were also created. The features then received MVA post-processing (mean subtraction, variance normalization, and ARMA filtering) [14]. MVA post-processing has been shown to give strong results on Aurora 3.0; therefore, our baseline results are already fairly good on this corpus [14, 15].

In all experiments the state observation (labeled $\mathcal{O}^s$) uses all 39 features, and its distribution is modeled as a 16 component Gaussian mixture model trained by maximizing the likelihood using EM. The baseline system is an HMM using only $\mathcal{O}^s$ and can be seen in Figure 2(a) [2]. Whole word models are used with 16 states per word, plus 3 states for a silence word, plus 1 state for short pause.

4. NEW FOCUSED MODELS

We evaluated a number of models that focus acoustic transition information directly on an ASR system's transition events. This section describes them all in detail.

The first new model, seen in Figure 2(b), is called the *Word Transition* model. It has an observation (labeled $\mathcal{O}^{wt}$) conditioned on word and word transition. $\mathcal{O}^{wt}$ uses only the 13 delta and 13 double delta features, and the model scores these features using only a single Gaussian component. This gives 26 (for Danish and German) or 24 (for Finnish and Spanish) additional single component 26 dimensional Gaussians. The $\mathcal{O}^{wt}$ Gaussians are also trained using maximum likelihood, but during their training the $\mathcal{O}^s$ Gaussians are initialized to the parameters that were learned for the baseline model and are held fixed. The transition probabilities, $p(\mathcal{P}^{tr}|\mathcal{P})$, however, are allowed to change while the transition Gaussians are training. This allows the new transition distributions to influence $p(\mathcal{P}^{tr}|\mathcal{P})$. In initial experiments this training method performed better than allowing the baseline parameters to change while the parameters for the additional Gaussians are training.

The next model is called *Word Plus Next Word* and is shown in Figure 2(c). When there is no word transition, $\mathcal{O}^{wt}$ is conditioned only on the current word. When there is a word transition, there are separate models dependent on the class of the next word. More precisely, for each word there is a model for transitioning from the word to silence, from the word to any other word (all grouped into one class), and from the word to a short pause. Silence and short pause are only allowed to transition into a word, so they have one model apiece. This is implemented in the graph using a backward time link from $\mathcal{W}_{t+1}$ to $\mathcal{W}_t$. This model has a total of 35 (for Danish and German) or 32 (for Finnish and Spanish) Gaussian components not in the baseline system.

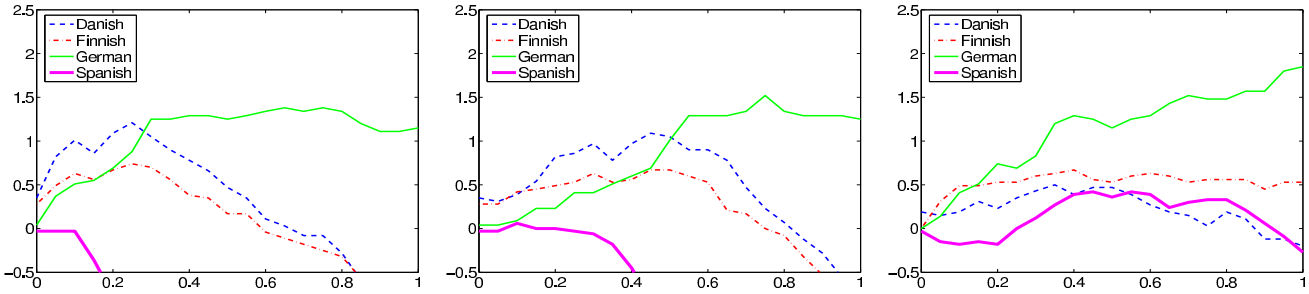
The third model is known as the *State Transition* model and is shown in Figure 2(d). This model contains an observation $\mathcal{O}^{st}$ containing the 13 delta and 13 double-delta features and uses a 26 dimensional single component Gaussian that is trained in the same way as $\mathcal{O}^{wt}$. In *State Transition* $\mathcal{O}^{st}$ is conditioned on the state and state transition, rather than on the word and word transition. This adds 360 (Danish and German) or 328 (Finnish and Spanish) components. This requires more parameters than the word transition graph but has the ability to influence within word transitions in addition to word segmentation.

State Plus Next Word is the next model and is shown in Figure 2(e). When there is no state transition or a within word transition $\mathcal{O}^{st}$ is conditioned on the current state and the state transition. When there is a transition out of a word the model works in an analogous fashion to *Word Plus Next Word*. For each word there is a model for transitioning from the word to silence, from the word to any other word, from the word to a short pause, and one model for a transition out of silence and another for a transition out of short pause. This adds 382 or 348 components.

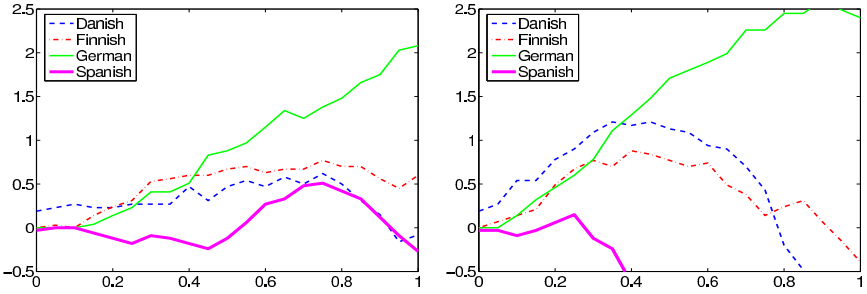
Finally, *Combined* puts together the observations from both the *Word Plus Next Word* model and the *State Plus Next Word* model. The Gaussian parameters that were trained separately for *Word Plus Next Word* and *State Plus Next Word* are used directly in *Combined* with no additional training. This gives a total of 417 (Danish and German) or 380 (Finnish and Spanish) additional components. Only one set of transition probabilities, $p(\mathcal{P}^{tr}|\mathcal{P})$, is needed to decode this model, and they are taken from *State Plus Next Word*.

5. RESULTS

We evaluate the aforementioned models on the highly mismatched task of the four languages in Aurora 3.0. In each of the models, the Gaussian observation scores need to be scaled (in an analogous manner to the acoustic scale factor used widely in LVCSR systems). This is because the two feature streams use different numbers of components and have different dimensionalities, and also because the scale can be used to control the degree of influence the observation has in deciding the result. In these experiments the scale of $\mathcal{O}^s$ is kept constant at 1, and the scale of either $\mathcal{O}^{wt}$ or $\mathcal{O}^{st}$ was tested over a range of values. Both observations were scaled to 1 (i.e. no scaling) during training. The Aurora 3.0 corpus does not provide development test sets, so a scale that works across all four data sets is crucial to indicate that the technique can be generalized rather than requiring tuning for a particular task. Although a development set would have been desirable, the recordings for the four languages were created by independent working groups under different noise conditions and the results are given for the case where there is a mismatch of noise conditions and microphones between training and testing. Figure 3 plots the absolute improvement over the baseline versus the scale.



(a) *Word Transition*, scale factor of $\mathcal{O}^{wt}$ (b) *Word Plus Next Word*, scale factor of $\mathcal{O}^{wt}$ (c) *State Transition*, scale factor of $\mathcal{O}^{st}$



(d) *State Plus Next Word*, scale factor of $\mathcal{O}^{st}$ (e) *Combined*, scale factor of $\mathcal{O}^{wt}$ and $\mathcal{O}^{st}$

Fig. 3. The models were decoded with an exponential scaling factor on the transition evidence feature stream. The scaling exponent is on the x axis and the absolute improvement over baseline is the y axis. Note that on Figures (a), (b), and (e) Spanish quickly falls below the bottom of the chart.

The single scale for each experiment was chosen based on the sum of the accuracy score for each language. The word recognition accuracies for each experiment at the chosen point is given in Table 1.

The *Word Transition* model shows considerable improvement over the baseline on Danish, French, and German but was not able to perform above the baseline on Spanish. *Word Plus Next Word* improves the curve on German and gives the other three language better performance over the range of scale values, but there is no point that improves the overall accuracy versus *Word Transition*. *State Transition* gives much improvement on German and Spanish, and Finnish performs over a larger range of scales. Danish does not do as well on the *State Transition* experiments as compared to the *Word Transition* experiments, but it is still above the baseline. *State Plus Next Word* gives a small improvement over *State Transition* for all four languages. It is interesting that on the two state transition graphs German was able to beat its baseline by 2.5 points, but only by using large scales (near 2). Scales this large on the other languages perform poorly. It is also notable that when considering only Danish, Finnish, and German the *Combined* model performed better than either *Word Plus Next Word* or *State Plus Next Word* alone. Unfortunately, as in *Word Plus Next Word*, Spanish does not do any better than the baseline.

One might wonder why *Word Transition* and *Word Plus Next Word* failed to show improvement on Spanish. One theory is that the final "s" found in three of the Spanish digits caused problems for these models. The "s" sound found elsewhere in the digits or in the noise might be prompting spurious word transitions. As evidence for this, compared to the baseline using a large scale value (0.8) on *Word Transition* gave 3.8 times as many insertions of the word "seis" and 2.1 times as many words mistranslated as "seis". The word "dos" had 2.3 times as many insertions and "tres" had 1.9 times as many insertions. No other word had both an order of magnitude increase and absolute increase of greater than 5 for a type of mistake. This theory is difficult to prove conclusively, though, and does not directly account for the entire dip in performance at high scale values.

6. CONCLUSION

Acoustic information for predicting word and state transitions was added to five graphical models at the part of the model where it was thought to most likely benefit ASR performance. The two models that conditioned on the State Transition variable were able to improve on the baseline for all four languages using a common scaling factor. The two models that conditioned on the Word Transition and

Table 1. Word accuracy scores at the best scaling points. The total accuracy is an average of the four individual scores. The first number in the # Parameters column is for Danish and German, the second number is for Finnish and Spanish.

Model	Scale	Danish	Finnish	German	Spanish	Total	# Parameters
Reference		31.90	75.41	74.28	42.23	55.96	
Baseline		80.53	91.10	88.81	90.71	87.79	$2.27, 2.07 \times 10^5$
Word Transition	0.1	81.54	91.73	89.32	90.68	88.32	$2.29, 2.08 \times 10^5$
Word Plus Next Word	0.3	81.50	91.73	89.22	90.65	88.28	$2.29, 2.09 \times 10^5$
State Transition	0.4	80.92	91.77	90.10	91.10	88.47	$2.46, 2.24 \times 10^5$
State Plus Next Word	**0.75**	**81.15**	**91.87**	**90.19**	**91.22**	**88.61**	$2.47, 2.25 \times 10^5$
Combined	0.35	81.74	91.80	89.92	90.47	88.48	$2.49, 2.27 \times 10^5$

the combined model showed improvements on three of the languages but failed to improve on the fourth. In the three cases where the combined system gave improvement it performed better than the individual models that it was composed of. Overall, we have shown that acoustic information can be focused and integrated into a variety of specific points in an ASR system, not just at the phone or state conditioned Gaussian mixture, and that this general approach can be quite beneficial. We plan in the future to combine the MVSE features (Mean and Variance of Spectral Entropy) defined in [3] with the approaches given here to hopefully further improve performance. We also plan to employ other forms of acoustic feature that could more beneficially indicate transition and/or speaking rate.

7. REFERENCES

[1] J. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, September 2005.

[2] J. Bilmes, G. Zweig, and et. al., "Discriminatively structured dynamic graphical models for speech recognition," in *In Final Report: JHU 2001 Summer Workshop*, 2001.

[3] A. Subramanya, J. Bilmes, and C. Chen, "Focused word segmentation for ASR," in *9th European Conf. on Speech Communication and Technology (Eurospeech)*, 2005.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1981, vol. 29, pp. 254 – 272.

[5] K. Elenius and M. Blomberg, "Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1982, vol. 7, pp. 535–538.

[6] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 34, no. 1, pp. 52–59, February 1986.

[7] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.* IEEE, 1990, pp. 857–860.

[8] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[9] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, and A.E. Rosenberg, "Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1991.

[10] J.G. Wilpon, C.-H. Lee, and L.R. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1991.

[11] N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky, "Stochastic perceptual auditory-event-based models for speech recognition," *Intl. Conf. on Spoken Language Proc.*, pp. 1943–1946, September 1994.

[12] J. Bilmes, N. Morgan, S.-L. Wu, and H. Bourlard, "Stochastic perceptual speech models with durational dependence," *Intl. Conf. on Spoken Language Proc.*, November 1996.

[13] J. Bilmes, *The GMTK Documentation.*

[14] C. Chen, K. Filali, and J. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases," in *Intl. Conf. on Spoken Language Proc.*, 2002.

[15] C. Chen, J. Bilmes, and D. Ellis, "Speech feature smoothing for robust ASR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, March 2005.