On the Exploitation of Hidden Markov Models and Linear Dynamic Models in a Hybrid Decoder Architecture for Continuous Speech Recognition

Volker Leutnant, Reinhold Haeb-Umbach
Department of Communications Engineering, University of Paderborn, Germany
{leutnant,haeb}@nt.uni-paderborn.de

Abstract

Linear dynamic models (LDMs) have been shown to be a viable alternative to hidden MARKOV models (HMMs) on small-vocabulary recognition tasks, such as phone classification. In this paper we investigate various statistical model combination approaches for a hybrid HMM-LDM recognizer, resulting in a phone classification performance that outperforms the best individual classifier. Further, we report on continuous speech recognition experiments on the AURORA4 corpus, where the model combination is carried out on wordgraph rescoring. While the hybrid system improves the HMM system in the case of monophone HMMs, the performance of the triphone HMM model could not be improved by monophone LDMs, asking for the need to introduce context-dependency also in the LDM model inventory.

Index Terms: speech recognition, hybrid decoder architecture, acoustic modeling, linear dynamic models

1. Introduction

Traditionally, automatic speech recognition systems are based on hidden MARKOV models (HMMs) with Gaussian mixtures modeling the state conditioned feature vector distributions. The power of modeling the speech feature trajectory by combining multimodal and multivariate Gaussians with an underlying hidden, discrete-valued state process and the inherent assumption of statistical independence between features in adjacent analysis frames once a HMM state is given has rendered this approach both effective and efficient. However, it's the conditional independence assumption that is commonly cited to be the major shortcoming of this prevailing acoustic modeling approach [1, 2].

Segment or trajectory models such as the linear dynamic models (LDMs) have been proposed to overcome this weakness [1, 2, 3, 4]. Linear dynamic models encounter these shortcoming by introducing a hidden, linear, autoregressive and continuous-valued state process underlying the observed features. Although performance of this modeling approach on phone classification tasks where phone boundaries are explicitly given has been found to be superior to that of an equivalent static model (1-state monophone HMM with unimodal full covariance Gaussian emission density), it falls short on performance compared to established acoustic modeling techniques, yet – both in phone classification with known phone boundaries and in unconstrained recognition of phones [5]. Further, application of LDMs to continuous speech recognition results in very challenging decoding operations, where approximations have to be introduced to make it computational tractable.

Irrespective of the preceding findings and issues it has been found that LDMs and HMMs have complementary modeling strenghts and that a combination of the two may result in improved recognition accuracies [5, 6]. While the HMM is capable of modeling long-term temporal dependencies, the strength of the LDM lies in the direct modeling of statistical dependencies between consecutive feature vectors.

In this paper we first review phone classification experiments on the TIMIT database. The results obtained by various statistical model combination approaches motivate the application of both LDMs and HMMs in a hybrid decoder architecture for continuous speech recognition. We show how this hybrid decoding is realized by rescoring wordgraphs and demonstrate by experiments that the combined system outperforms the best individual one in the case of context-independent models. Since, however, the performance of the triphone HMM decoder could not be improved by application of the LDMs, we conclude the paper with a discussion on how to extend LDMs to realize context-dependent segment models.

2. Hidden MARKOV Models

The hidden MARKOV model is the most popular approach to model the observed speech features and is capable of modeling long-term temporal dependencies. By introducing a hidden, discrete-valued state process underlying the observation process, the likelihood of a sequence of feature vectors $\mathbf{y}_{t_{s(n)}}^{t_{e(n)}} = \left(\mathbf{y}_{t_{s(n)}}, \mathbf{y}_{t_{s(n)}+1}, .., \mathbf{y}_{t_{e(n)}}\right)$, $\mathbf{y}_\tau \in \mathbb{R}^d$, starting at discrete time instance $t_{s(n)}$ and ending at discrete time instance $t_{e(n)}$ given a hypothesized word $\omega_n \in [\Omega_1, .., \Omega_K]$, where n denotes the position of the word within the sentence,. is given by

$$p(\mathbf{y}_{t_{s(n)}}^{t_{e(n)}}|\omega_n) \approx \max_{q_{t_{s(n)}}^{t_{e(n)}}} \prod_{\tau=t_{s(n)}}^{t_{e(n)}} p(\mathbf{y}_\tau|q_\tau,\omega_n) P(q_\tau|q_{\tau-1},\omega_n). \quad (1)$$

The maximization in the above VITERBI-approximation has to be carried out over all possible sequences $q_{t_{s(n)}}^{t_{e(n)}}$ of hidden states making up the word under consideration, subject to the state transition probabilities $P(q_\tau|q_{\tau-1},\omega_k)$ within this word. The state-conditioned feature vector distribution $p(\mathbf{y}_\tau|q_\tau,\omega_n)$ is usually modeled as a mixture of I Gaussians

$$p(\mathbf{y}_\tau|q_\tau = j, \omega_n = \Omega_k) = \sum_{i=1}^{I} c_{i,j,k} \mathcal{N}\left(\mathbf{y}_\tau; \boldsymbol{\mu}_{i,j,k}, \boldsymbol{\Sigma}_{i,j,k}\right), \quad (2)$$

with mixture weights $c_{i,j,k}$, means $\boldsymbol{\mu}_{i,j,k}$ and covariances $\boldsymbol{\Sigma}_{i,j,k}$. In large vocabulary speech recognition, the word HMMs are usually obtained from concatenating HMMs based on subword units such as mono- or triphones, both trained under the expectation maximization (EM) framework.

3. Linear Dynamic Models

Linear dynamic models have been proposed as an alternative acoustic model for phone classification [3] and recognition [7]. The LDM system is based on a hidden, linear, autoregressive, continuous-valued state process underlying the observation process. A linear measurement equation relates the hidden state to

the observation. Although conceptually LDMs and HMMs both utilize a Markovian state space model, the continuity of the state space and the direct statistical dependencies between adjacent feature vectors render the computation of the likelihood of a sequence of feature vectors $\mathbf{y}_{t_{s(n)}}^{t_{e(n)}}$ given a hypothesized word ω_n nontrivial. However, assuming subword boundaries within words to be known, the likelihood of the words can be computed by switching the model parameters at subword boundaries. For a given subword unit $\upsilon_l(n) \in [\Upsilon_1, .., \Upsilon_M]$ (with l denoting the position of the subword unit within the word ω_n) starting at time instance $t_{s(n,l)}$ and ending at time instance $t_{e(n,l)}$, the likelihood can be computed as

$$p(y_{t_{s(n,l)}}^{t_{e(n,l)}}|\upsilon_l(n)) = \prod_{\tau=t_{s(n,l)}}^{t_{e(n,l)}} p(y_\tau|y_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n)), \quad (3)$$

with the subword unit $\upsilon_l(n)$ absorbing the dependency on the word ω_n. Conceptually, (3) is evaluated by introducing the continuous-valued state sequence $\mathbf{x}_{t_{s(n,l)}}^{t_{e(n,l)}}$, $\mathbf{x}_\tau \in \mathbb{R}^{d'}$, underlying the observation sequence $\mathbf{y}_{t_{s(n,l)}}^{t_{e(n,l)}}$ as

$$p(\mathbf{y}_{t_{s(n,l)}}^{t_{e(n,l)}}|\upsilon_l(n)) = \prod_{\tau=t_{s(n,l)}}^{t_{e(n,l)}} \int_{\mathbf{x}_\tau} p(\mathbf{y}_\tau|\mathbf{x}_\tau, \upsilon_l(n)) p(\mathbf{x}_\tau|\mathbf{y}_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n))\, d\mathbf{x}_\tau \quad (4)$$

and computing $p(\mathbf{x}_\tau|\mathbf{y}_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n))$ by recursively exploiting

$$p(\mathbf{x}_\tau|\mathbf{y}_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n)) = \int_{\mathbf{x}_{\tau-1}} p(\mathbf{x}_\tau|\mathbf{x}_{\tau-1}, \upsilon_l(n)) p(\mathbf{x}_{\tau-1}|\mathbf{y}_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n)) d\mathbf{x}_{\tau-1}, \quad (5)$$

$$p(\mathbf{x}_\tau|\mathbf{y}_{t_{s(n,l)}}^{\tau}, \upsilon_l(n)) \propto p(\mathbf{y}_\tau|\mathbf{x}_\tau, \upsilon_l(n)) p(\mathbf{x}_\tau|\mathbf{y}_{t_{s(n,l)}}^{\tau-1}, \upsilon_l(n)), \quad (6)$$

which can be solved analytically if state and measurement equation are linear and driven by (uncorrelated) Gaussian noises, resulting in the standard KALMAN filtering. The probability density functions completely describing the LDM are given by

$$p(\mathbf{x}_1|\upsilon_l(n)=\Upsilon_m) = \mathcal{N}(\mathbf{x}_1; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (7)$$

$$p(\mathbf{x}_\tau|\mathbf{x}_{\tau-1}, \upsilon_l(n)=\Upsilon_m) = \mathcal{N}(\mathbf{x}_\tau; \mathbf{F}_m\mathbf{x}_{\tau-1}+\mathbf{w}_m, \mathbf{D}_m), \quad (8)$$

$$p(\mathbf{y}_\tau|\mathbf{x}_\tau, \upsilon_l(n)=\Upsilon_m) = \mathcal{N}(\mathbf{y}_\tau; \mathbf{H}_m\mathbf{x}_\tau+\mathbf{v}_m, \mathbf{C}_m), \quad (9)$$

with $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{F}, \mathbf{w}, \mathbf{D}, \mathbf{H}, \mathbf{v}, \mathbf{C}\}_{m=1}^M$ denoting the means and covariances of the state priors, the state transition matrices, the means and covariances of the state noises, the measurement matrices and the means and covariances of the measurement noises, respectively. The likelihood of a sequence of subword units $\{\upsilon_l(n)\}_{l=1}^{L(n)}$ of word and pronunciation dependend length $L(n)$, eventually composing the word ω_n, can be computed in two ways, namely by following the *state-reset* (R) or the *state-passed* (P) approach, as denoted by FRANKEL [5].

In the *state-reset* approach any acoustic context dependency between successive subword units is disregarded and the state ans thus the KALMAN filter is forced to be reset at the beginning of a new subword unit. The likelihood of a sequence of subword units in the *state-reset* approach can thus be written as the product of the individual segment likelihoods

$$p^{(\mathrm{R})}(\mathbf{y}_{t_{s(n)}}^{t_{e(n)}}|\omega_n) \approx \prod_{l=1}^{L(n)} p(y_{t_{s(n,l)}}^{t_{e(n,l)}}|\upsilon_l(n)), \quad (10)$$

where $p(y_{t_{s(n,l)}}^{t_{e(n,l)}}|\upsilon_l(n))$ is given by (3). In contrast to the *state-reset* approach, the *state-passed* approach allows the state to be continuous over subword boundaries. The likelihood computation thus consists of a single KALMAN filter recursion over all features $\mathbf{y}_{t_{s(n)}}^{t_{e(n)}}$ associated with the hypothesized word ω_n. However, the filter's parameters are switching at subword boundaries. Introducing the discrete-valued indicator variable $\upsilon(\tau)$ specifying the subword unit and the LDM parameters to be used at time instance τ one obtains the likelihood of a hypothesized word as

$$p^{(\mathrm{P})}(y_{t_{s(n)}}^{t_{e(n)}}|\omega_n) = \prod_{\tau=t_{s(n)}}^{t_{e(n)}} p(y_\tau|y_{t_{s(n)}}^{\tau-1}, \upsilon(\tau)). \quad (11)$$

For both approaches, *state-reset* and *state-passed*, EM algorithms can be applied to train subword LDM parameters.

4. Statistical Model Combination

Statistical model combination aims at combining the modeling power of two or more models in the hope that the combined model will be better than the individual ones [8]. Applying this paradigm to the acoustic modeling at hand thus asks for rules to combine the traditional HMMs with the recently proposed LDMs, both modeling subword units. In general, statistical model combination of multiple acoustic models for speech recognition can happen either on the "likelihood level" or the "subword unit posterior level". In the first case, the final likelihood of the hypothesized word under consideration is a function of the individual likelihoods, whereas in the latter case the final posterior probability of the hypothesized word under consideration is computed as a function of the posterior probabilities of the individual models.

5. Phone Classification on TIMIT

Initial experiments on statistical model combination have been carried out on the TIMIT phone classification task. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions for each utterance [9]. The decimated version of the data with a sampling rate of 8 kHz was employed. The phonetic transcriptions have been used to train and evaluate 61 context-independent LDMs and HMMs on the standard $d = 39$ dimensional MFCC+Δ+Δ^2 feature vectors obtained by coding the speech data with the ETSI standard front-end [10]. In the evaluation phase, LDM and HMM scores for all hypothesized phones of a given utterance have been stored in a wordgraph and baseline LDM and HMM classification results have been determined by applying the VITERBI-search on the wordgraph. The following model combination techniques have been examined:

- HMM/LDM: always select either the HMM or the LDM;
- model combination based on likelihoods:
  - Product LH: the support for a phone is the exponentially weighted product of the individual likelihoods;
- model combination based on phone posterior probabilities calculated on wordgraphs [11]:
  - Product PP: the support for a phone is the exponentially weighted product of the individual posterior probabilities;
  - Inverse Entropy: the support for a phone is the weighted sum of the individual posterior probabilities with weights inversely proportional to the entropy of the acoustic models [12];
  - Entropy-based DS: DEMPSTER-SHAFER model combination [12]; weights of the ignorance models are based on the entropy of the acoustic models;

The results listed in Table 1 have been obtained using the *state-reset* approach with LDMs of state dimensionality $d' = 12$ and HMMs with a 3-state linear topology and a mixture of 20 diagonal covariance Gaussians modeling the state conditioned feature vector distributions. A log-Gaussian duration model and an unsmoothed phone bigram language model have been trained on the same corpus as the LDMs and HMMs and are applied, too (see [6] for details). All results are reported on a collapsed phone set of cardinality 39. If explicit weighting is involved, weights for the HMM and the LDM are denoted by w_{HMM} and w_{LDM}, respectively. Although the HMM incorporates many more parameters per phone (4747) compared to the LDM (1611) and although the HMM system achieves a considerably better classification accuracy than the LDM system, the hybrid HMM-LDM classifier always outperforms the best individual one. The increases in classification accuracy can be found to be significant and are apparent over all model combination methods examined. However, the more sophisticated model combination methods based on phone posterior probabilities (e.g. Entropy-based DEMPSTER-SHAFER model combination) yield, if any, only marginally better results than the "simple" exponentially weighted product of likelihoods with equally weighted models. Choosing the weights in the product rules different from the default values of $w_{\text{HMM}} = w_{\text{LDM}} = 0.5$ further improves the classification accuracy. Optimal values of ($w_{\text{HMM}} = 0.7, w_{\text{LDM}} = 0.3$) for the exponentially weighted product of likelihoods and ($w_{\text{HMM}} = 0.75, w_{\text{LDM}} = 0.25$) for the exponentially weighted product of phone posterior probabilities have been determined during a development phase leading to a classification accuracy of $77.64\,\%$ and $77.81\,\%$, respectively.

Table 1: *Classification results for the proposed model combination approaches on the TIMIT phone classification task*

model combination approach studied	classification accuracy [%]
HMM	76.96
LDM	73.28
Product LH ($w_{\text{HMM}} = w_{\text{LDM}} = 0.5$)	77.46
Product LH ($w_{\text{HMM}} = 0.7, w_{\text{LDM}} = 0.3$)	77.64
Product PP ($w_{\text{HMM}} = w_{\text{LDM}} = 0.5$)	77.62
Product PP ($w_{\text{HMM}} = 0.75, w_{\text{LDM}} = 0.25$)	77.81
Inverse Entropy	77.42
Entropy-based DS	77.45

6. Speech Recognition on AURORA4

Due to the effectiveness and the simplicity, the exponentially weighted product of likelihoods will be applied to the speech recognition experiments following. However, as noted in section 3, the computation of the likelihood for a given word hypothesis ω_n under the LDM is not trivial. The exploitation of direct statistical dependencies between consecutive feature vectors $\mathbf{y}_{t_{s(n)}}^{t_{e(n)}}$ by the LDMs asks for the consideration of all possible model (subword unit) histories from the beginning of a hypothesized word $t_{s(n)}$ to its end $t_{e(n)}$ (which, in general, is hypothesized, too). This requires a separate KALMAN filter to be run for each hypothesized start time of the corresponding subword unit in the *state-reset* case. If the *state-passed* approach is considered, the computational complexity is further increased, since separate KALMAN filters have to be run for each possible sequence of model histories.

To avoid the exploration of all possible paths for a given word hypothesis and the involved computation, this paper exploits the alignment capabilities of the HMM recognizer to facilitate the application of the LDMs to the problem of continuous speech recognition. That is, a wordgraph storing alternative word hypotheses is constructed for each utterance under consideration [13]. Besides the word hypotheses with start and end times, the state sequence corresponding to the best path through the HMM state trellis spanned for each hypothesis is kept, too. With this limited representation of the search space at hand, the computational burden associated with the use of LDMs for continuous speech recognition is thus reduced to rescoring of time-aligned sequences of subword units in the wordgraphs by either applying the *state-reset* or the *state-passed* approach. However, since FRANKEL [5] found the *state-passed* approach to perform worse than the *state-reset* approach on the phone classification task, the focus in this paper is on the *state-reset* approach.

The experiments were performed on the AURORA4 database. The AURORA4 test database consists of the Wall-street Journal Nov'92 evaluation test set to which noise at varying SNR levels and of varying type has been added [14]. In this paper, the official AURORA4 selection test set comprising 166 utterances recorded with a Sennheiser microphone and decimated to $8\,\text{kHz}$ is used. Beside the clean data of the AURORA4 selection test set, six versions of the test set with artificially added noises at randomly chosen SNR conditions between $5\,\text{dB}$ and $15\,\text{dB}$ are examined. The results for the *clean* test set as well as the average results over all noisy sets, both obtained by using a bigram language model for the 5000-word vocabulary, will be presented. Training of triphone and monophone HMMs with a 3-state linear topology has been carried out on clean training data, coded into the standard $d = 39$ dimensional MFCC+Δ+Δ^2 feature vectors by using the ETSI standard front-end [10]. While the monophone HMM is based on unimodal diagonal covariance Gaussians, the triphone HMM bases on mixtures of 10 diagonal covariance Gaussians. Monophone LDMs are trained on the triphone-aligned feature vectors obtained by operating the recognizer in forced-alignment mode using the already trained triphone HMMs on the clean training data. In doing so, 43 LDMs (42 monophone models + 1 model for silence) of state dimension $d' = 12$ remain to be trained. The triphone based HMM recognizer is further used to construct a wordgraph for each utterance [13]. During rescoring of word hypotheses by the monophone LDMs, the monophone alignment for each word hypothesis in the wordgraph is obtained by using the triphone alignment while dropping the triphone context. LDM rescoring follows the state-reset approach presented in section 3. Rescoring with monophone HMMs solely uses the hypotheses' start and end times with word HMMs being constructed from monophone HMMs by utilizing the dictionary, followed by searching for the optimal path through the trellis spanned by the HMM states.

Finally, recognition results are obtained by applying the VITERBI-search on the wordgraph using either the monophone HMM acoustic scores, the monophone LDM acoustic scores or the combined acoustic scores following the exponentially weighted product of likelihoods model combination approach. The weights for the HMM and the LDM, respectively, have been set to the ones found to give the best performance on the TIMIT phone classification task, i.e. ($w_{\text{HMM}} = 0.7, w_{\text{LDM}} = 0.3$). Resulting (E)rrors, (D)eletions, (S)ubstitutions and (I)nsertions are given in Tables 2 and 3. Note that in general the error rates are fairly large, which is due to the fact that monophones rather than triphones are used and, for the noisy data, no attempt has been made to compensate for the noise. Hence, the trends reported in the following are consistent over all data sets.

Table 2: *Recognition results for rescoring the wordgraph with monophone HMMs or monophone LDMs*

test	monophone HMM				LDM			
set	S [%]	D [%]	I [%]	E [%]	S [%]	D [%]	I [%]	E [%]
clean	18.93	2.25	6.37	27.55	20.33	4.05	7.55	31.93
noisy	50.22	11.04	11.92	73.17	48.04	18.68	7.06	73.78

Table 3: *Recognition results for rescoring the wordgraph by combining monophone HMM and LDM likelihoods*

test	monophone HMM + LDM			
set	S [%]	D [%]	I [%]	E [%]
clean	17.38	2.32	6.48	26.19
noisy	42.33	14.32	9.34	70.98

7. Discussion

As can be seen from Table 2, the recognition performance of the monophone HMMs and the monophone LDMs differ significantly in the percentage of insertions and deletions, with the LDM system being prone to deletion errors, as also observed by FRANKEL [7]. However, no attempt has been taken to balance the number of insertions and deletions (e.g. by tuning the word insertion penalty). Looking at the error rates, the monophone HMM system gives better results than the monophone LDM system. Results of moving from the individual systems to a hybrid decoder architecture by combining the likelihoods of the monophone HMMs and the monophone LDMs in the exponentially weighted product approach are given in Table 3. The combined acoustic likelihoods finally yield consistently better recognition results than the best individual system for all noise types. Monophone linear dynamic models with the *state-reset* approach to compute the likelihoods under the LDMs are thus able to aid recognition of continuous speech with an acoustic model based on monophone HMMs.

When context-dependent triphones are used, the error rate of the HMM recognizer improves to $16.80\,\%$, which is about half of the error rate of the (monophone) LDM recognizer. Due to this large performance difference, the combination of triphone HMM likelihoods with the LDM likelihoods did no longer lead to an improvement in error rate, as shown in Table 4. For the LDM to be beneficial, its error rates need to

Table 4: *Recognition results for rescoring the wordgraph with triphone HMMs and their combination with LDMs*

test	triphone HMM				triphone HMM + LDM			
set	S [%]	D [%]	I [%]	E [%]	S [%]	D [%]	I [%]	E [%]
clean	10.20	1.10	5.49	16.80	10.76	1.33	5.49	17.57
noisy	35.77	10.17	11.02	56.96	34.77	14.12	8.46	57.32

be brought into the vicinity of the performance of the triphone HMM recognizer. The most promising way to achieve this is to incorporate context-dependency also into the LDM system. However, the *state-passed* approach may not be appropriate to introduce sufficient context dependency. An interesting option would be to employ multiple (switching) linear dynamic models (SLDMs) to model a subword unit. SLDMs have been shown to be an appropriate model of speech dynamics in the context of feature enhancement for robust speech recognition [15, 16, 17]. Finding ways to keep the number of model parameters from increasing beyond train- and tractable dimensions (e.g. parameter tying, model adaptation) is just one potential challenge of the multiple model approach and, as the use of SLDMs for modeling phonetic context in general, will be left for future research.

8. Conclusions

In this paper the combination of linear dynamic models and hidden MARKOV models based on Gaussian mixtures as an acoustic model for continuous speech recognition is considered. With significant improvements on a preliminary phone classification task, the combination of HMM and LDM acoustic scores has been applied to continuous speech recognition on the AURORA4 corpus, where the model combination is carried applied to wordgraph rescoring. While the hybrid system has been found to improve the HMM system in the case of monophone HMMs, the performance of the triphone HMM model could not be improved by monophone LDMs, asking for further exploration of incorporating and modeling phonetic context also in an LDM system.

9. Acknowledgement

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/6-1.

10. References

[1] M. Ostendorf, V. Digalakis, and O. Kimball, "From hmm's to segment models: a unified view of stochastic modeling for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 5, pp. 360 –378, September 1996.

[2] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153 – 173, 2007.

[3] V. Digalakis, J. Rohlicek, and M. Ostendorf, "Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[4] M. Russell and W. Holmes, "Linear trajectory segmental hmms," *Signal Processing Letters, IEEE*, vol. 4, no. 3, pp. 72 –74, March 1997.

[5] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 246–256, January 2007.

[6] V. Leutnant and R. Haeb-Umbach, "Options for modelling temporal statistical dependencies in an acoustic model for asr," in *Proceedings of DAGA 2010*. DAGA, March 2010.

[7] J. Frankel, "Linear dynamic models for automatic speech recognition," Ph.D. dissertation, University of Edinburgh, Edinburgh, UK, 2003.

[8] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, July 2004.

[9] W. M. Fisher, G. R. Doddington, Goudie-Marshall, and K. M., "The darpa speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[10] ETSI ES 201 108, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI Std. ES 201 108, Rev. 1.1.3, September 2003.

[11] F. Wessel, "Word posterior probabilities for large vocabulary continuous speech recognition," Ph.D. dissertation, RWTH Aachen, Aachen, Germany, 2002.

[12] F. Valente, "Multi-stream speech recognition based on dempster-shafer combination rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, 2010.

[13] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, 1997.

[14] H.-G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task au/417/02," STQ AURORA DSR WORKING GROUP, Tech. Rep., November 2002.

[15] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, A. Acero, Ed., vol. 1, 2004, pp. I–953–6 vol.1.

[16] V. Leutnant and R. Haeb-Umbach, "An analytic derivation of a phase sensitive observation model for noise robust speech recognition," in *Interspeech 2009*, Interspeech. Interspeech, September 2009, pp. 2395–2398.

[17] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," in *Proceedings of Interspeech 2009*, Interspeech. Interspeech, September 2009, pp. 1231–1234.