MULTI-FRAME FACTORISATION FOR LONG-SPAN ACOUSTIC MODELLING

Liang Lu and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

{liang.lu, s.renals}@ed.ac.uk

ABSTRACT

Acoustic models based on Gaussian mixture models (GMMs) typically use short span acoustic feature inputs. This does not capture long-term temporal information from speech owing to the conditional independence assumption of hidden Markov models. In this paper, we present an implicit approach that approximates the joint distribution of long span features by product of factorized models, in contrast to deep neural networks (DNNs) that model feature correlations directly. The approach is applicable to a broad range of acoustic models. We present experiments using GMM and probabilistic linear discriminant analysis (PLDA) based models on Switchboard, observing consistent word error rate reductions.

Index Terms— Acoustic modelling, long span features, multi-frame factorisation

1. INTRODUCTION

Hidden Markov model (HMM) based speech recognition [1] typically uses acoustic feature vectors that are extracted from a short temporal window of context at a fixed frame rate. These acoustic feature vectors are assumed conditionally independent given the HMM state sequence, which allows efficient acoustic model training and decoding. However, this has been viewed as one of major weaknesses of HMM-based acoustic models [2], because there is significant temporal dependence between neighbouring acoustic frames – a dependency that cannot be captured efficiently by conventional HMM-based acoustic models.

Conventional HMM systems, using Gaussian mixture model (GMM) output distributions, do incorporate some acoustic context, for example through the use of dynamic features [3] or through the projection of multiple acoustic frames using linear discriminant analysis [4]. These approaches work well in practice, although they violate the dependence assumptions of the HMM. Since the additional dynamic features are arrived at through a linear transform of the original feature vector sequence, it is possible to interpret the resultant acoustic models as defining a probability over the complete observation sequence. This may be considered as a different model, referred to as the trajectory HMM, and the training and decoding algorithms can be derived accordingly [5].

Segment-based acoustic models [6, 7] aim to model subword units – such as phones or sub-phones – directly. These approaches model the joint distribution of variable-length observation sequences (with or without the conditional independence assumption) and the segment models are used as the basic building blocks, rather than HMM states. More recently segmental conditional random fields (CRFs) [8] have been proposed. Segmental CRFs differ from the previously proposed models through the use of multi-scale detectors to extract segmental-level features, using different CRF feature functions which are integrated to predict the state sequence.

Neural network (NN) acoustic models have achieved considerable success using concatenated acoustic vectors from wider windows of context to take advantage of the long temporal acoustic information for classification. These approaches have resulted in improvements in accuracy in both hybrid NN/HMM systems [9, 10, 11, 12, 13] and in GMM-based systems using tandem or bottleneck features obtained from NNs trained for context-independent or context-dependent phone classification [14, 15, 16]. In the case of tandem or bottleneck features, the HMM independence assumptions are again violated (and accuracy is again improved).

In this paper, we investigate generative acoustic models which directly model multiframe context. Modelling the joint distribution of concatenated features is normally prohibitive for generative models owing to the high dimensionality, and applying explicit dependency between multiple frames – such as in the autoregressive HMM [1] and linear dynamic systems [17] – makes efficient inference difficult. We consider a simple approach which factorizes the joint distribution of multiple frames by a product of the probabilities of the individual frames. This can be viewed as a type of segment model that uses a fixed length segmentation, thus avoiding the need to estimate latent segmentation variables. This approach is very general and is applicable to a wide range of acoustic models. In this paper, probabilistic linear discriminant analysis (PLDA) [18] (cf. Section 4) is used as an example. Our experiments on Switchboard demonstrate that the approach results in reductions in word error rate (WER).

Funded by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

2. HIDDEN MARKOV MODEL AND LONG-SPAN ACOUSTIC MODELLING

2.1. Hidden Markov Model

Given a sequence of acoustic observations of length T, $\mathbf{O} = (\mathbf{o}_1, \cdots, \mathbf{o}_T)$, the optimal word sequence $\hat{\mathbf{W}}$ is obtained by the *maximum a posteriori* (MAP) decoding rule as

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{O}|\mathcal{M}, \mathbf{W}) P(\mathbf{W}) \tag{1}$$

where $\mathcal{M}$ denotes the acoustic model parameters, and $P(\mathbf{W})$ is the prior probability obtained from a language model. Using HMMs, the acoustic likelihood score is computed as

$$p(\mathbf{O}|\mathcal{M}, \mathbf{W}) = \sum_{\mathbf{q} \in \Phi_{\mathbf{W}}} p(\mathbf{O}|\mathbf{q}, \mathcal{M}) P(\mathbf{q}) \tag{2}$$

where $\mathbf{q} = (q_0, q_1, \cdots, q_T, q_{T+1})$ denotes an HMM state sequence corresponding to the observations, and $\Phi_{\mathbf{W}}$ denotes the set of all the possible HMM state sequences of $\mathbf{W}$. q_0 and q_{T+1} are non-emitting entry and exit states that are not considered in the following for simplicity. Equation (2) can be solved efficiently if we apply the *first-order Markov* and *conditional independence* assumptions:

$$P(\mathbf{q}) \approx \prod_{t=1}^{T} P(q_t|q_{t-1}), \quad p(\mathbf{O}|\mathbf{q}, \mathcal{M}) \approx \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{q}, \mathcal{M}) \tag{3}$$

Given that, we can rewrite equation (2) as

$$p(\mathbf{O}|\mathcal{M}, \mathbf{W}) \approx \sum_{\mathbf{q} \in \Phi_{\mathbf{W}}} \prod_{t=1}^{T} p(\mathbf{o}_t|q_t, \mathcal{M}) P(q_t|q_{t-1}) \tag{4}$$

which is behind the HMM-based acoustic models.

2.2. Long-span acoustic modelling

To capture long term dependence, one way is to address the conditional independence assumption of HMMs directly. For instance, some Bayesian models introduce an explicit dependency between observations in equation (3), e.g.

$$p(\mathbf{O}|\mathbf{q}, \mathcal{M}) \approx \prod_{t=2}^{T} p(\mathbf{o}_t|\mathbf{o}_{t-1}, \mathbf{q}, \mathcal{M}) p(\mathbf{o}_1|\mathbf{q}, \mathcal{M}) \tag{5}$$

Typical examples include autoregressive HMM [1] and switching linear dynamic systems [17]. However, these models are usually expensive to train, and do not work well in practice.

Another approach is to use long span features directly by concatenating the observations of certain context size without formulating the internal dependence explicitly. If we define

$$\mathbf{y}_t := (\mathbf{o}_{t-k}, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_{t+k}) \tag{6}$$

and if we do not consider the effect of feature overlapping when computing the state transition probability, this model may be represented as

$$p(\mathbf{O}|\mathcal{M}, \mathbf{W}) \approx \sum_{\mathbf{q} \in \Phi_{\mathbf{W}}} \prod_{t=1}^{T} p(\mathbf{y}_t|q_t, \mathcal{M}) P(q_t|q_{t-1}) \tag{7}$$

where $2k+1$ is the context size. Though it is not very precise theoretically, this approach works well for DNN-based acoustic models [11, 12]. In practice, the short span features $\{\mathbf{o}_{t+n}, n \in [-k, k]\}$ may overlap with each other in the time domain, and if $\mathbf{o}_{t+n}$ is beyond the utterance length, the empty elements can be padded by the first or last frames.

Using long span features such as $\mathbf{y}_t$ is impractical for GMM-based acoustic models due to the expansion of model size and diagonalisation of covariance. DNNs, however, do not have this difficulty, and the scaled likelihood can be computed efficiently as

$$p(\mathbf{y}_t|q_t, \mathcal{M}) \propto P(q_t|\mathbf{y}_t, \mathcal{M}) / P(q_t) \tag{8}$$

Previously, we studied the use of PLDA to estimate the joint probability distribution $p(\mathbf{y}_t|q_t, \mathcal{M})$ directly which is able capture feature correlations and is more scalable to higher dimensional features [18]. However, this approach does not work well for very high dimensional features. In this paper, we investigate an implicit approach which approximates this joint distribution by product of factorized models.

3. MULTI-FRAME FACTORISATION

The idea of factorizing the distribution of $p(\mathbf{y}_t|q_t, \mathcal{M})$ is not new. For instance, the *semi-parametric trajectory* model [19], rewrites the joint distribution as

$$p(\mathbf{y}_t|q_t, \mathcal{M}) = p(\boldsymbol{\tau}_t|\mathbf{o}_t, q_t, \mathcal{M}_\tau) p(\mathbf{o}_t|q_t, \mathcal{M}_o) \tag{9}$$

where $\boldsymbol{\tau}_t := (\mathbf{o}_{t-k}, \ldots, \mathbf{o}_{t-1}, \mathbf{o}_{t+1}, \ldots, \mathbf{o}_{t+k})$, and we have used $\mathcal{M}_\tau$, and $\mathcal{M}_o$ to show that the two distributions may be modelled by different sets of model parameters. In [19], the model is simplified by dropping the dependency on $\mathbf{o}_t$ as

$$p(\boldsymbol{\tau}_t|\mathbf{o}_t, q_t, \mathcal{M}_\tau) \approx p(\boldsymbol{\tau}_t|q_t, \mathcal{M}_\tau) \tag{10}$$

In this case the factorisation is performed on each Gaussian component for a GMM-based acoustic model; here we do not introduce a Gaussian component index, in order to clarify the presentation of the basic idea.

In this paper, we present a multi-frame model which factorizes the joint probability distribution into a product of probabilities of each individual frames. This can be written as

$$p(\mathbf{y}_t|q_t, \mathcal{M}) \approx \prod_{n=-k}^{k} p^{\gamma_n}(\mathbf{o}_{t+n}|q_t, \mathcal{M}_n) \tag{11}$$

$$s.t. \quad \sum_n \gamma_n = 1 \tag{12}$$

where $\mathcal{M}_n$ denotes the parameters that model the distribution of the individual frames $\mathbf{o}_{t+n}$, and γ_n is the scaling factor which normalises the distribution and compensates for any factorisation error. Compared to standard HMMs (equation (7)), this approach predicts the states by averaging scores from a wider context. Note that the factorisation can be performed at the state level or the Gaussian component level [19].

This approach does not rely on a particular type of acoustic model, and a similar approach may be applied to DNNs [20], where the state posterior probability is obtained as

$$p(q_t|\mathbf{y}_t, \mathcal{M}) \approx \prod_{n=-k}^{k} p^{\gamma_n}(q_t|\mathbf{o}_{t+n}, \mathcal{M}_n) \qquad (13)$$

where $\gamma_n = \frac{1}{2k+1}$, and here $\mathbf{o}_{t-n}$ itself is a long span feature vector, as normally used in DNNs. The authors report considerably improved accuracy using this approach [20]. In this paper, we show that the multi-frame factorisation approach is also applicable to generative models. Note that in (11) some model parameters for each short span features $\mathcal{M}_n$ may be shared: for instance, the hidden layers of DNNs may be shared across all the factorized models as in [20] , while for structured generative models such as the subspace GMM (SGMM) [21] and PLDA [22], the state-independent model parameters can be tied across all the factorized models, which is similar to multilingual training.

4. PLDA-BASED ACOUSTIC MODEL

As discussed before, multi-frame factorisation is applicable to general types of acoustic models including DNN hybrid models, e.g. [20]. In this paper, we apply the approach to a generative model based on PLDA. Our motivation is that the PLDA acoustic model is more flexible with respect to higher dimensional features compared to GMMs, allowing us to compare the results of modelling the joint distribution $p(\mathbf{y}_t|q_t, \mathcal{M})$ directly to multi-frame factorisation.

The basic idea behind PLDA acoustic models is that the distribution over acoustic feature vectors $\mathbf{y}_t \in \mathbb{R}^d$ from the j-th HMM state at time t (i.e. $q_t = j$) is expressed as [18]:

$$\mathbf{y}_t|j = \mathbf{U}\mathbf{x}_{jt} + \mathbf{G}\mathbf{z}_j + \mathbf{b} + \epsilon_{jt}, \quad \epsilon_{jt} \sim \mathcal{N}(\mathbf{0}, \Lambda)\,. \qquad (14)$$

$\mathbf{z}_j \in \mathbb{R}^q$ is the state variable shared by the whole set of acoustic frames generated by the j-th state and $\mathbf{x}_{jt} \in \mathbb{R}^p$ is the frame variable which explains the per-frame variability. Usually, the dimensionality of these two latent variables is smaller than that of the feature vector $\mathbf{y}_t$, i.e. $p, q \leq d$. $\mathbf{U} \in \mathbb{R}^{d\times p}$ and $\mathbf{G} \in \mathbb{R}^{d\times q}$ are two low rank matrices which span the subspaces to capture the major variations for $\mathbf{x}_{jt}$ and $\mathbf{z}_j$ respectively. $\mathbf{b} \in \mathbb{R}^d$ denotes the bias and $\epsilon_{jt} \in \mathbb{R}^d$ is the residual noise which is assumed to be Gaussian with zero mean and diagonal covariance. By marginalising out ϵ_{jt} and $\mathbf{x}_{jt}$, we obtain the following likelihood function:

$$p(\mathbf{y}_t|j) = \mathcal{N}(\mathbf{y}_t; \mathbf{G}\mathbf{z}_j + \mathbf{b}, \mathbf{U}\mathbf{U}^T + \Lambda) \qquad (15)$$

To increase the modelling capacity, we presented a tied PLDA mixture model [22], which computes the state likelihood as

$$p(\mathbf{y}_t|j) = \sum_{mk} w_{jkm}\mathcal{N}\left(\mathbf{y}_t; \mathbf{G}_m\mathbf{z}_{jk} + \mathbf{b}_m, \mathbf{U}_m\mathbf{U}_m^T + \Lambda_m\right)$$

where k denotes the sub-state index, and $\mathbf{z}_{jk}$ is the sub-state variable, c_{jk} is the sub-state weight, π_{jm} is the component weight which is shared for all the sub-state models, and $w_{jkm} = c_{jk} \times \pi_{jm}$. This model is closely related to the SGMM [21], and more details can be found in [22].

5. EXPERIMENTS AND DISCUSSION

We performed experiments using the Switchboard corpus [23]. The Hub-5 Eval 2000 data [24] is used as the test set, which contains the Switchboard (SWB) and CallHome (CHM) evaluation subsets. The experiments were performed using the Kaldi speech recognition toolkit [25], and we have used GMM- and tied PLDA-based acoustic models. In the following experiments, we used maximum likelihood estimation without speaker adaptation. The pronunciation lexicon was obtained from the Mississippi State transcriptions [26] and a trigram language model was used for decoding.

5.1. Baseline systems

Table 1 shows the WERs of the baseline systems, trained using about 33 hours of Switchboard training data. The number of tied HMM states is around 2,400 for each of the acoustic models shown in this table. The GMM system has about 30,000 Gaussian components. In the PLDA system, the state and frame variables are both 40-dimensional, irrespective of the acoustic feature dimensions. More details about this setup can be found in [22]. Since the PLDA acoustic model is more flexible with respect to feature vectors dimension, we evaluated the effect of long span features by splicing 13-dimension MFCC_0 of different context size. Table 1 shows that these features can improve the accuracy of PLDA systems, however, no further improvement was achieved when the feature length is longer than MFCC_0($\pm$3), i.e. splicing 3 left/right frames with the current one. Previously, we used spliced MFCC_0_Δ_$\Delta\Delta$ as features but did not obtain better results [18]. As a comparison, we also show the WERs of GMM systems using the same long span features but with feature space linear discriminant analysis (LDA) which reduces the feature dimensionality to 40 and followed by semi-tied covariance matrix (STC) modelling [27]. Table 1 shows that the PLDA systems consistently outperform their GMM counterparts.

5.2. Results of using multi-frame factorisation

Based on our previous results, we use MFCC_0($\pm$3) as the basic feature unit $\mathbf{o}_t$ in the following experiments to evaluate the multi-frame factorisation using PLDA. Before presenting

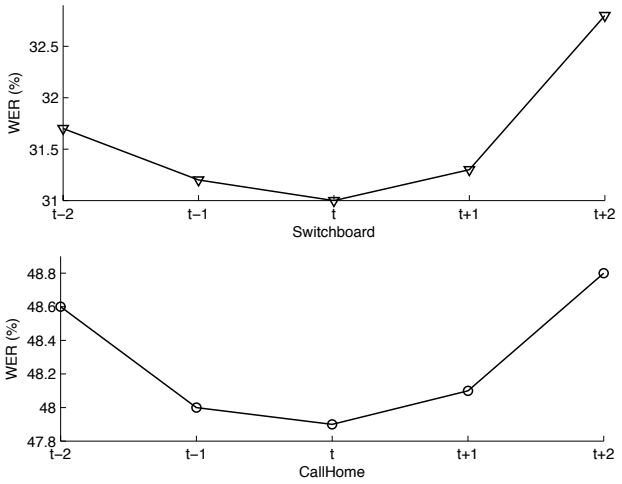

Fig. 1. WERs of systems $p(\mathbf{o}_{t+n}|q_t, \mathcal{M}), n \in [-2, 2]$ on the Switchboard and CallHome evaluation set.

Table 1. WER(%) of GMM and PLDA baseline systems using 33 hours of training data.

System	Feature	CHM	SWB	Avg
GMM	MFCC_0+Δ+$\Delta\Delta$	54.0	36.6	45.4
GMM	MFCC_0($\pm$2)+LDA_STC	54.4	34.4	43.7
GMM	MFCC_0($\pm$3)+LDA_STC	50.6	33.5	42.2
GMM	MFCC_0($\pm$4)+LDA_STC	50.7	33.3	42.1
GMM	MFCC_0($\pm$5)+LDA_STC	50.9	34.1	42.4
PLDA	MFCC_0 ($\pm$2)	48.6	31.9	40.4
PLDA	MFCC_0 ($\pm$3)	47.9	31.0	39.5
PLDA	MFCC_0 ($\pm$4)	47.5	31.2	39.4
PLDA	MFCC_0 ($\pm$5)	48.7	32.2	40.6

these results, we first evaluate the systems that use individual feature frames, i.e. the state likelihood score is computed as $p(\mathbf{o}_{t+n}|q_t, \mathcal{M}_n), n \in [-k, k]$. These systems were trained in the usual way except that given the alignment, we used $\mathbf{o}_{t+n}$ instead of $\mathbf{o}_t$ to accumulate statistics for q_t to train model parameters. The aim is to study the effect of mismatch between the alignments and feature inputs, and the correlations between the consecutive frames. In addition, these systems are also baselines for the multi-frame factorisation system since it averages the scores from those systems. These results are shown in Figure 1. We observe that shifting the feature input by 1 time step results in a marginal performance degradation, otherwise the systems obtained much worse results. It will be interesting to see if this trend holds for DNNs.

We then used multi-frame factorisation to integrate the factorized models using (11), which is referred to as `MF-State` in Table 2 because the factorisation is performed at the state level. In this work, the scaling factors γ_n were simply set to be $\frac{1}{2k+1}$. However, their values can either be tuned manually or learned by Bayesian optimisation approach [28]. As a comparison, we also performed system combination using minimum Bayes risk (MBR) decoding by combining the word lattices from each sub-system [29], which is referred as `MF-Sequence`. Since MBR decoding itself may perform better than MAP, we show both results for baseline systems for strict comparison. Table 2 shows that although the sub-systems in Figure 1 are worse than the baseline, both state and sequence level integration result in lower WER. The `MF-State` system does not achieve better results when $k = 2$ because the sub-systems trained on $\mathbf{o}_{t-2}$ and $\mathbf{o}_{t+2}$ are very poor, and using equal scaling factors γ_n is not optimal. However, this system still works much better than that without factorisation which corresponding to PLDA–MFCC_0($\pm$5) system in Table 1. These results indicate that the factorisation approach can work better than directly estimate the joint distribution of long span features. A similar trend was observed when using 300 hours of training data as shown by Table 3. In future, we shall investigate using wider feature context and factorisation approach (13) for DNNs.

Table 2. Results of using multi-frame factorisation for PLDA systems using different feature context $n \in [-k, k]$.

System	k	Decoding	CHM	SWB	Avg
Baseline	0	MAP	47.9	31.0	39.5
MF-State	1	MAP	47.3	30.3	38.8
MF-State	2	MAP	48.4	30.8	39.7
Baseline	0	MBR	47.2	30.3	38.9
MF-State	1	MBR	46.6	29.7	38.3
MF-State	2	MBR	47.9	30.3	39.3
MF-Sequence	1	MBR	46.2	29.6	37.9
MF-Sequence	2	MBR	**45.8**	**29.4**	**37.7**

Table 3. Results of using 300 hours of training data.

System	k	Decoding	CHM	SWB	Avg
Baseline	0	MAP	40.8	25.1	33.1
MF-State	1	MAP	40.2	24.6	32.5
MF-State	2	MAP	41.3	25.5	33.5
Baseline	0	MBR	40.2	24.6	32.6
MF-State	1	MBR	39.7	24.3	32.2
MF-State	2	MBR	40.8	25.2	33.2
MF-Sequence	1	MBR	39.2	24.2	31.8
MF-Sequence	2	MBR	**39.1**	**24.1**	**31.7**

6. CONCLUSION

In this paper, we study multi-frame factorisation for long-time acoustic modelling where the distribution of long-span acoustic features is factorized into a product of short-time models. Compared to approaches that apply explicit dependency between observation for long temporal modelling, this approach is more efficient in model training, and is applicable to a wide range of acoustic models. Experiments on Switchboard demonstrate that this approach can improve speech recognition accuracy. The factorisation approach presented in this paper is very simple. In the future, we shall investigate approaches to reduce the factorisation error and to approximate the joint distribution more precisely.

7. REFERENCES

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

[2] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *Proc. ASRU*. IEEE, 2011, pp. 71–76.

[3] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 52–59, 1986.

[4] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. IEEE ICASSP*, vol. 2, 2000, pp. 1129–1132.

[5] 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 and Language*, vol. 21, pp. 153–173, 2007.

[6] M. Ostendorf, V. V. Digalakis, and O. A. 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, 1996.

[7] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech & Language*, vol. 17, no. 2, pp. 137–152, 2003.

[8] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*. IEEE, 2009, pp. 152–157.

[9] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer, 1994, vol. 247.

[10] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[11] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[12] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[14] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*. IEEE, 2000, pp. 1635–1638.

[15] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos, "Pushing the envelope - aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[16] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE ICASSP*, vol. 4, 2007, pp. 757–760.

[17] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. ICASSP*, vol. 1. IEEE, 2004, pp. I–953.

[18] L. Lu and S. Renals, "Probabilistic linear discriminant analysis for acoustic modelling," *IEEE Signal Processing Letters*, 2014.

[19] S. Liu and K. C. Sim, "Temporally varying weight regression: A semi-parametric trajectory model for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 151–160, 2014.

[20] N. Jaitly, V. Vanhoucke, and G. Hinton, "Autoregressive product of multi-frame predictions can improve the accuracy of hybrid models," in *Proc. INTERSPEECH*, 2014.

[21] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[22] L. Lu and S. Renals, "Tied probabilistic linear discriminant analysis for speech recognition," *arXiv:1411.0895 [cs.CL]*, 2014.

[23] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*. IEEE, 1992, pp. 517–520.

[24] C. Cieri, D. Miller, and K. Walker, "Research methodologies, observations and outcomes in (conversational) speech data collection," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 206–211.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovský, G. Semmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[26] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of SWITCHBOARD," in *Proc. ICSLP*, 1998.

[27] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[28] S. Watanabe and J. Le Roux, "Black box optimization for automatic speech recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 3256–3260.

[29] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.