WEAK TOP-DOWN CONSTRAINTS FOR UNSUPERVISED ACOUSTIC MODEL TRAINING

Aren Jansen, Samuel Thomas, Hynek Hermansky

Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD USA

ABSTRACT

Typical supervised acoustic model training relies on strong top-down constraints provided by dynamic programming alignment of the input observations to phonetic sequences derived from orthographic word transcripts and pronunciation dictionaries. This paper investigates a much weaker form of top-down supervision for use in place of transcripts and dictionaries in the zero resource setting. Our proposed constraints, which can be produced using recent spoken term discovery systems, come in the form of pairs of isolated word examples that share the same unknown type. For each pair, we perform a dynamic programming alignment of the acoustic observations of the two constituent examples, generating an inventory of cross-speaker frame pairs that each provide evidence that the same subword unit model should account for them. We find these weak top-down constraints are capable of improving model speaker independence by up to 57% relative over bottom-up training alone.

Index Terms— speaker independent acoustic models, unsupervised training, spectral clustering, top-down constraints

1. INTRODUCTION

The human speech recognition ability is acquired over several years of exposure to caregivers who provide top-down supervision via linguistic or visual context, enabling specialization to our native language. Automatic speech recognition technology implements equivalent top-down training mechanisms that rely on large, lexically transcribed corpora and pronunciation dictionaries that relate the acoustic feature space to a phonetic inventory for subword modeling. In the zero resource setting, we are without the requisite training materials and language-specific knowledge to provide this standard form of top-down supervision. However, the past five years have seen several efforts in developing scalable automatic approaches to discovering repeated words and phrases in large speech corpora using nothing but the raw feature vectors as input [1, 2, 3, 4, 5, 6]. While these spoken term discovery methods are useful in their own right for tasks like keyword search [7, 8] and topic discovery [9, 10], they also automatically uncover valuable information regarding the lexical structure of the language.

In this paper, we are interested in using automatically discovered lexical units to aid in unsupervised training of speaker independent phonetic acoustic models. At the core of our proposed approach is the assumption that word level patterns are easier to identify across speakers than frame-level ones. Figure 1(a) shows spectrograms for two examples of the phoneme /iy/, one spoken by a female and the other a male. In isolation, the acoustic pattern similarity is not immediately obvious. Figure 1(b) shows spectrograms for two examples of the word encyclopedia spoken by the same male and female. At the word level, the visual similarities are much more striking: we see similar temporal alternation between voiced/unvoiced segments and low/high frequency energy, and we notice similar formant contours even though the absolute formant frequencies are different. In fact, the vowels of Figure 1(a) are none other than the first /iy/ segments from the two word examples of Figure 1(b). Our conclusion is that fully unsupervised learning in speech should ideally begin at the word level, where we have a better chance of discovering repeated patterns across speakers. However, once we identify repetition at the word level, we still need a mechanism for channeling that information into the construction of a speaker independent subword acoustic model for downstream recognition tasks.

In the present study, we take as given a collection of word segment pairs of the same unknown type and thus assume each also has the same underlying subword unit sequence. Thus, if we perform a dynamic time warping (DTW) alignment of the constituent feature vectors extracted for each word segment pair, we can expect corresponding frames to map to the same subword unit (or at least the same posterior distribution across the subword units). Here, monotonicity of the warping function will force phonetically equivalent but acoustically dissimilar portions of the word segments to align. We can channel these frame-level constraints to cluster Gaussian components from a large universal background model of speech, where each cluster defines a Gaussian mixture model that corresponds to some speaker independent subword unit. In a series of proof-of-concept experiments, where we use true word example pairs as a noise-free proxy for the output of a term discovery system, we demonstrate that the proposed weak constraint mechanism can substantially improve speaker independence relative to bottom-up training alone. Moreover, we find that the number of word level constraints required to achieve a significant improvement is surprisingly limited.

2. RELATED WORK

For more than a decade, the term unsupervised training in acoustic modeling has referred to using lightly supervised models to generate noisy transcriptions for unannotated speech, which are fed back for subsequent retraining [11, 12]. In the past five years, however, truly unsupervised subword acoustic model training has been attempted using various bottom-up strategies, including Gaussian mixture-based universal background models [7], successive state splitting algorithms for hidden Markov models (HMM) [13], traditional estimation of subword HMMs [14], discriminative clustering objectives [15], and non-parametric Bayesian estimation of HMMs [16]. These approaches are united by the fact that none impose top-down constraints of any kind and thus exhibit limited speaker independence properties.

Two more recent efforts have proposed using automatically discovered words to constrain bottom-up unsupervised training procedures. The first [10] uses an initial bottom-up model to tokenize the speech; word discovery is subsequently performed using this

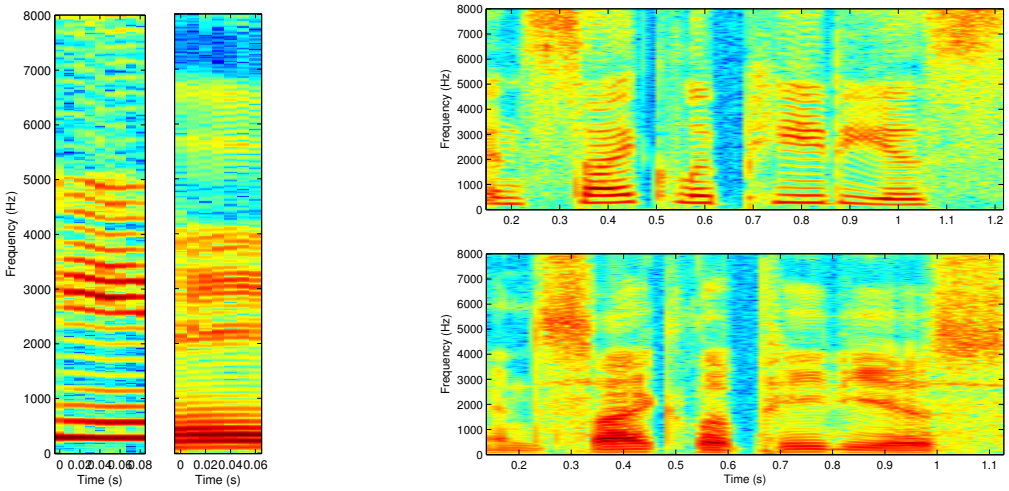

Fig. 1. Cross-speaker word repetition is easier to identify than phone repetition. Shown are log power spectrograms for isolated phoneme (a) and word (b) segments, spoken by a male and female speaker.

noisy 1-best tokenization, an order of operations that can obscure the speaker independence of whole word acoustic patterns. The direct precursor [17] to this paper uses automatically discovered clusters of word examples to train whole word HMMs with Gaussian mixture emission densities and then clusters HMM states across the word models to produce context independent subword unit models. While successful, the main limitation of that approach was that several examples of each automatically discovered word type were necessary to construct the whole word model, meaning if a word was only uttered a few times it could not contribute. Moreover, only speech contained in the repeated word segments could contribute to model parameter estimation. Our proposed method circumvents these limitations by (i) using the entire speech collection to estimate a large universal background model and (ii) using individual repeated word segment pairs to partition UBM Gaussians into subword unit GMMs.

3. UNSUPERVISED TRAINING ALGORITHM

Like its predecessor in [17], our present strategy for learning speaker independent subword acoustic models is to discover repeated word-level patterns in the raw acoustic stream and use them to constrain unsupervised clustering in the acoustic space. Thus, we require a large collection of untranscribed speech for unsupervised training. Given this data, the proposed training procedure consists of four steps (see Figure 2): (1) train a large GMM-based universal background model (UBM) using a large sample of in-domain audio; (2) run a spoken term discovery system, such as that presented in [6], across the speech collection to produce a collection of word segment pairs and compute UBM posteriorgrams for each segment; (3) perform a DTW alignment of the acoustic frames of each word segment pair and use the corresponding posteriorgram frame pairs to construct a similarity matrix over UBM Gaussian components; and (4) using spectral clustering, partition the UBM Gaussian components and use each subset to define a subword unit GMM. Our focus in this paper is the efficacy of the top-down constraints, so we will take as given the collection of same word segment pairs of the form produced by a spoken term discovery system. Recent improvements in term discovery scalability [6] support obtaining an arbitrarily large number of word segment pairs. However, we evaluate performance as a function of the number of these pairs in order to understand better how many will be required for decent performance.

3.1. The Universal Background Model

Given a large collection of untranscribed speech audio, the first step is to compute a short-time feature vector time series representation (e.g. PLP or MFCC) of the form $X = x_1 x_2 \ldots x_T$ where $x_t \in \mathbb{R}^d$, and train in a completely bottom-up fashion a large universal background model for all speech (and silence) content using maximum likelihood estimation. We define our UBM to be a Gaussian mixture model with C components of the form

$$P(x) = \sum_{c=1}^{C} \alpha_c \mathcal{N}(x; \mu_c, \Sigma_c), \tag{1}$$

where $\{\alpha_c\}$ are the mixture weights and $\mathcal{N}(x; \mu_c, \Sigma_c)$ is the d-dimensional multivariate normal for the c-th UBM component with with mean μ_c and covariance matrix Σ_c. For C sufficiently large, each Gaussian component will cover a region of the acoustic feature space that corresponds to some speaker- and/or context-dependent subword unit. Nevertheless, the UBM is a soft vector quantization that provides sort of crude acoustic model that imposes a categorical structure on the acoustic space for various downstream tasks, e.g. [7, 18, 8]. Given its usage in previous efforts [7, 16], the UBM will serve as our baseline for our speaker independence evaluation in Section 4.

3.2. Partitioning UBM Components

Our top-down constraints come in the form of a collection of N repeated word segment pairs, which we denote $\{(X_i, Y_i)\}_{i=1}^{N}$, where $X_i = x_1 x_2 \ldots x_{A_i}$ for $x_t \in \mathbb{R}^d$ and $Y_i = y_1 y_2 \ldots y_{B_i}$ for $y_t \in \mathbb{R}^d$ are the acoustic feature vectors for the i-th pair. We can convert each word segment pair into a collection of frame-level correspondences by performing a dynamic time warping alignment [19] using cosine distance as the frame-level metric. Taken together, the N word segment pairs produce F frame pairs of the form $\mathcal{F} = \{(x_i, y_i)\}_{i=1}^{F}$,

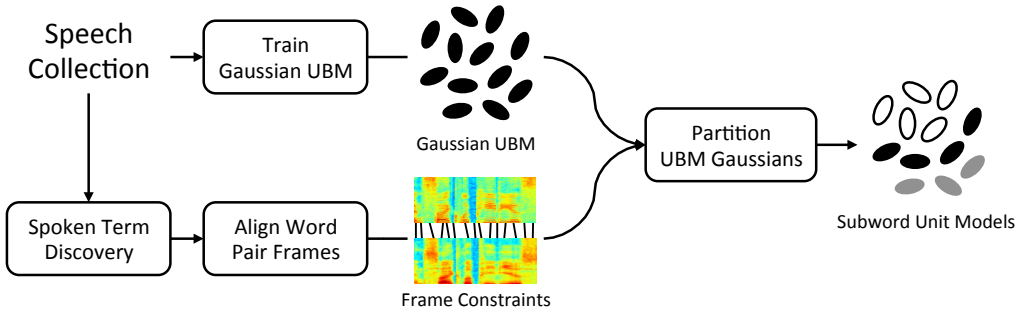

Fig. 2. Training algorithm schematic.

where $x_i, y_i \in \mathbb{R}^d$. Note that time warping tolerated by the alignment means that while each frame pair is unique, each individual frame can occur in multiple frame pairs (limited by the natural variation in phone duration).

The next step is to relate the frame level pairs $\mathcal{F}$ to the components of the UBM for subsequent partitioning. At a high level, each speaker independent subword unit will consist of a subset of the Gaussian components of the UBM that tend to simultaneously activate for frame pairs in $\mathcal{F}$. We can obtain the requisite UBM component co-occurrence statistics as follows. For each $(x_i, y_i) \in \mathcal{F}$, we can compute the posterior distribution over the UBM components for a given acoustic frame x by

$$P(c|x) = \frac{\mathcal{N}(x; \mu_c, \Sigma_c)}{\sum_{c'=1}^{C} \mathcal{N}(x; \mu_{c'}, \Sigma_{c'})}, \quad (2)$$

where we have assumed a uniform component prior by discarding the GMM mixing weights $\{\alpha_c\}$. We can then compute an aggregate $C \times C$ (soft) co-occurrence matrix between UBM components by

$$S(c_1, c_2) = \frac{\sum_{i=1}^{F} P(c_1|x_i) P(c_2|y_i)}{\left[\sum_{i=1}^{F} P(c_1|x_i)\right] \left[\sum_{i=1}^{F} P(c_2|y_i)\right]}, \quad (3)$$

which has been normalized by the expected counts of each UBM component. The goal then is to partition the set of UBM components such that pairs of UBM components that have high values in S fall into the same subset.

Having demonstrated success in a similar setting [17], we use spectral clustering to derive the partition into K subsets as follows First, the co-occurrence matrix S is used to define a weighted undirected graph with C vertices, each corresponding to a single Gaussian component of the UBM. Each matrix element S_{ij} specifies the edge weight between the vertices corresponding to the i-th and j-th component. Unlike more common agglomerative techniques, spectral clustering attempts not to just group vertices that are directly similar, but also those that are connected by paths of high similarity. Given a desired number of clusters K, we implement the spectral clustering variant of [20]:

1. Compute the unnormalized graph Laplacian $L = D - S$, where D is the diagonal matrix with elements $D_{ii} = \sum_j S_{ij}$, the degree of the i-th vertex.
2. Solve the generalized eigenvalue problem $Lv = \lambda Dv$, for the first K eigenvectors $\{v_1, \ldots, v_K\}$, where each $v_i \in \mathbb{R}^C$.
3. Representing the i-th vertex (and thus the i-th UBM component) by its graph spectrum $u_i = \langle v_1[i], v_2[i], \ldots, v_K[i] \rangle \in \mathbb{R}^K$, perform K-means clustering of the points $\{u_1, u_2, \ldots, u_C\}$.

The K-way clustering of vertices corresponds to a K-way partition of the Gaussian components. Each subset of Gaussian components itself then defines a Gaussian mixture model, where we assume a uniform mixture weight on each component. In this way, we have transformed the speaker dependent UBM into a collection of K Gaussian mixture models, each corresponding to a subword unit that we will demonstrate below in Section 4 exhibits substantially improved consistency across speaker relative to the UBM.

4. EXPERIMENTS

We perform several experiments to evaluate the speaker independence enabled by the proposed weak top-down constraint mechanism. We use a training set of cepstral mean and variance normalized perceptual linear prediction (PLP) features [21] corresponding to 40 hours of speech (180 conversations) from the Switchboard corpus of English conversational telephone speech. Our implementation details for the evaluation are as follows:

(a) **Building the universal background model:** GMM-based UBMs are trained bottom-up using maximum likelihood (ML) estimation. Starting with a single Gaussian component, training proceeds by interleaving Gaussian splitting and expectation-maximization re-estimation steps, which are performed until the desired number C of mixture components is reached. Diagonal covariance matrices are used in all cases. The GMM models are trained on only speech regions of the training corpus as identified by an neural network based speech activity detector [22]. We train baseline UBMs for $C = 50, 100, 150, 200$ and 1024 components, where we assume each component corresponds to some speaker- and/or context-dependent subword unit.

(b) **Deriving frame-level correspondences:** In practice, our word segment pairs can be generated using a scalable spoken term discovery algorithm such as that described in [6]. For the present evaluation, we forgo automatic discovery to limit extrinsic error sources and instead extract word segment pairs from a forced-alignment of the transcripts for the 40 hour train set. Restricting ourselves to word segments of at least 0.5 seconds in duration and 5 characters as text (the approximate bounds necessary for reliable term discovery [6]), we are left with nearly $N = 100,000$ same-type word segment pairs. Using DTW alignment of the PLP features for each pair, we generate approximately $F = 7$ million frame-level correspondences. We evaluate performance using all 100,000 word pairs, as well as for random subsamples of sizes $N =$ 10,000, 1,000, and 100.

(c) **Partitioning UBM components:** Next, we compute posteriorgrams using the 1024-component UBM for each word segment

pair according to Eqn. 2. Given the frame-level correspondences and the UBM posterior distribution for each frame, we populate the aggregate soft co-occurrence matrix using Eqn. 3. Since $C = 1024$, this produces a 1024×1024 similarity matrix, which defines the graph edge weights for spectral clustering of the 1024 Gaussian components into K subsets. We consider clusterings into $K = 50, 100, 150$, and 200 classes. Analogous to the UBMs constructed in Step (a), we treat each subset of the partition as model for some subword unit.

After training both baseline UBMs and weakly constrained models, we proceed to generate posteriorgrams for evaluation. For the baseline UBMs trained in Step (a), frame level posteriors are computed using Eqn. 2. Posteriors for the weakly constrained models are generated by collapsing the 1024-component UBM posteriors according to the partitions learned in Step (c) for various values of K.

Our goal is to evaluate these acoustic model posteriorgrams for suitability in downstream multi-speaker search and recognition tasks. In [23], a procedure was proposed to evaluate the quality of speech representations in the absence of a strict phonetic interpretability. It uses a large collection of presegmented word examples and computes the DTW distance between all example pairs, quantifying how well it can differentiate between pairs of same and different type. Average precision (AP), defined as the area under the precision-recall curve, is used to summarize performance over all DTW threshold operating points. Since the word examples are drawn from a wide range of speakers, this AP metric measures representational consistency across speaker. Moreover, in the case of supervised acoustic models, it was demonstrated the AP is perfectly correlated with phone recognition accuracy. This makes the AP metric an appealing proxy when evaluating unsupervised acoustic models, since calculation of phone recognition accuracy becomes impossible. Our instantiation uses 11K word examples drawn from a portion of the Switchboard corpus that was distinct from the 40 hour training set described above. These 11k examples result in 60.7M word pairs of which 96K are same type pairs across a wide variety of speakers (only 3K same-type pairs are from the same speaker). Cosine distance is the frame-level metric for the PLP baseline, while a more meaningful symmetrized Kullback-Leibler divergence is used for acoustic model posteriors (both supervised and unsupervised).

Table 1 summarizes the performance of 4 different feature representations on the evaluation set. These include raw acoustic features (PLP with mean/variance normalization), posteriors from the UBMs with various numbers of components, posteriors from the weakly constrained models with various number of unit clusters, and phoneme posteriors from fully supervised neural network ("English NN") acoustic models [23] trained on both 10 and 100 hours. Although the baseline UBM posteriors are only marginally better than raw acoustic features, the weakly constrained model posteriors show significant gains over bottom-up training alone. With comparable posterior dimension (i.e. for $K = C$), we find our weak constraints provide relative improvements in average precision ranging from 21%-57%. Since average precision has been demonstrated to be very well correlated with phone recognition accuracy for supervised models [23], these relative improvements can be thought of in accuracy terms as well. While there is still a substantial gap between the weak and strong top-down supervision, our proposed method bridges 37% of the gap between the UBM and 10-hour NN performance when C and K are both 100. Note that optimal performance occurs when C is approximately 2-3 times the number of phones, indicating the discovery of units akin to phone tristates. As such, integrating temporal continuity constraints via more sophisticated bottom-up models (e.g. [16, 13]) is a logical next step.

Table 1. Average precision (AP) performance on the word matching task for the baselines (unsupervised UBM and supervised English neural network) and proposed method, considering various values of C for the baseline UBM experiments and units target units K using the proposed top-down constraints. The best unsupervised performance is highlighted in boldface.

Features	*AP*
PLP w/MVN	0.194
UBM, $C = 50$	0.151
UBM, $C = 100$	0.196
UBM, $C = 150$	0.207
UBM, $C = 200$	0.222
UBM, $C = 1024$	0.222
UBM-1024 + Constraints, $K = 50$	0.238
UBM-1024 + Constraints, $K = 100$	**0.286**
UBM-1024 + Constraints, $K = 150$	0.275
UBM-1024 + Constraints, $K = 200$	0.270
English NN, 10 hr	0.439
English NN, 100 hr	0.516

Table 2. Average precision (AP) performance on the word matching task for the proposed method ($K = 100$) as a function of various values of N and F, the number word-level and corresponding frame-level top-down constraints, respectively.

N	F	*AP*
10^5	7×10^6	0.286
10^4	7×10^5	0.284
10^3	7×10^4	0.266
10^2	7×10^3	0.206

In a second set of experiments, we vary the number of word-level constraints and, consequently, the number of frame-level constraints that we impose. Table 2 shows performance as the constraints are reduced by several orders of magnitude, while keeping the number of clustered components at the optimal value of 100. We find that with as few as 1,000 word pairs we can still observe a significant improvement over both the raw feature performance and the baseline UBM performance. To put these numbers in context, the term discovery system presented in [6] can easily process hundreds of hours of speech, producing millions of candidate word segment pairs. While false alarms will be present, this level of scalability permits confidence thresholds to be set as high as necessary and still recover sufficient word pairs to achieve the gains demonstrated here.

5. CONCLUSIONS

We have presented a new strategy for unsupervised learning of a subword acoustic model that tempers bottom-up EM training with weak top-down lexical constraints generated by unsupervised term discovery systems. In the absence of transcribed speech and pronunciation dictionaries, these constraints provide substantial speaker independence gains over bottom-up training alone. In future work, we will explore the complementarity of our top-down constraints with more sophisticated bottom-up modeling techniques, e.g. [16, 13, 24], to help bridge the remaining gap between fully supervised methods.

6. REFERENCES

[1] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE T-ASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[2] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *Interspeech*, 2007.

[3] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Interspeech*, 2009.

[4] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. of ICASSP*, 2010.

[5] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[6] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011.

[7] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *ASRU*, 2009.

[8] A. Jansen and B. Van Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. Interspeech*, 2012.

[9] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP*, 2010.

[10] T.J. Hazen, M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Topic modeling for spoken documents using only phonetic information," in *Proc. of the ASRU*, 2011.

[11] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proc. of ICASSP*, 2002.

[12] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labeled data," in *Proc. of ICASSP*, 2009.

[13] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic subword units," in *ACL-08: HLT*, 2008.

[14] M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised audio patterns discovery using HMM-based self-organized units," in *Proc. of Interspeech*, 2011.

[15] X. Anguera, "Speaker independent discriminant feature extraction for acoustic pattern matching," in *Proc. ICASSP*, 2012.

[16] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. ACL*, 2012.

[17] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Interspeech*, 2011.

[18] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without asr," in *Proc. EMNLP*. Association for Computational Linguistics, 2010, pp. 460–470.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, 1978.

[20] J. Shi and J. Malik, "Normalized cuts and image segmenation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[21] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[22] S. Ganapathy, S. Thomas, and H. Hermansky, "Static and dynamic modulation spectrum for speech recognition," in *Proc. of Interspeech*, 2009.

[23] M. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. of ICASSP*, 2011.

[24] A. Jansen, S. Thomas, and H. Hermansky, "Intrinsic spectral analysis for zero and high resource speech recognition," in *Proc. of Interspeech*, 2012.