INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Supervised Learning of Acoustic Models in a Zero Resource Setting to Improve DPGMM Clustering

Michael Heck, Sakriani Sakti, Satoshi Nakamura

Augmented Human Communication Laboratory,
Graduate School of Information Science,
Nara Institute of Science and Technology,
Nara, Japan

{michael-h,ssakti,s-nakamura}@is.naist.jp

Abstract

In this work we utilize a supervised acoustic model training pipeline without supervision to improve Dirichlet process Gaussian mixture model (DPGMM) based feature vector clustering. We exploit methods common in supervised acoustic modeling to unsupervisedly learn feature transformations for application to the input data prior to clustering. The idea is to automatically find mappings of feature vectors into sub-spaces that are more robust to channel, context and speaker variability. The need of labels for these techniques makes it difficult to use them in a zero resource setting. To overcome this issue we utilize a first iteration of DPGMM clustering to generate frame based class labels for the target data. The labels serve as basis for learning an acoustic model in the form of hidden Markov models (HMMs) using linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and speaker adaptive training (SAT). We show that the learned transformations lead to features that consistently outperform untransformed features on the ABX sound class discriminability task. We also demonstrate that the combination of multiple clustering runs is a suitable method to further enhance sound class discriminability.

Index Terms: acoustic unit discovery, Bayesian nonparametrics, Dirichlet process, feature transformation, Gibbs sampling, unsupervised linear discriminant analysis, zero resource

1. Introduction

In a *zero resource scenario*, large amounts of labeled training data, parallel data, and knowledge about the target language are unavailable for developing speech processing systems with supervised techniques. Albeit significant advances in developing methods for unsupervised learning, current speech processing technology is not yet capable to imitate the natural capacities of infants to robustly learn acoustic and language models in an unsupervised way. Specialized evaluations such as the zero resource speech challenge [1] address this demanding task.

Confronted with an unknown language, phonologists usually attempt to define a set of acoustic units to fully cover the underlying sound repertoire. Machine learning approaches to this task are pattern matching [2, 3] on raw audio data and unsupervised sound unit detection [4]. These techniques have been successfully applied to solve tasks such as spoken term detection [5], topic segmentation [6] or document classification [7].

Model complexity usually is not known a priori when dealing with new data sets. Bayesian models such as the Dirichlet process Gaussian mixture model (DPGMM) can automatically adjust the model complexity given some data and have already been successfully applied to speech processing tasks such as unsupervised lexical clustering [8]. Chen et al. [9] cluster standard MFCC speech features by inferring a DPGMM and demonstrate its suitability for automatic detection of sound classes in untranscribed data. Their work is the best-performing contribution to the zero resource speech challenge 2015 [1].

Speech processing systems typically utilize feature transformations to increase sound class discriminability. Linear discriminant analysis (LDA) [10] is a standard technique to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. Maximum likelihood linear transforms (MLLT) [11, 12] and feature-space maximum likelihood linear regression (fMLLR) [13, 14] are commonly used to de-correlate feature components and for speaker adaptation. Naturally, class discriminating properties are critical for clustering, and adaptive feature transformations can help dealing with speaker variability. However, methods such as LDA need class labels for estimating the feature transformations, making it difficult to use them in a zero resource setting where the classes and even their amount are unknown.

In previous work [15] we demonstrated that it is possible to learn LDA transformations on automatically generated labels, and that these transformations can be used to produce feature vectors that considerably improve clustering performance. There has been work that utilize k-means clustering to automatically obtain pseudo labels for LDA estimation [16, 17]. But unlike in these studies we were able to overcome the limitation of having to predefine the size of prospective label sets by utilizing the non-parametric DPGMM sampler for clustering.

In this work we improve the DPGMM clustering by utilizing multiple feature transformations that can be combined with the previously exploited LDA transformations. Labels that were automatically generated in a first pass of DPGMM clustering serve as basis for learning an acoustic model using LDA, MLLT and fMLLR in an entirely unsupervised fashion. We demonstrate that each feature transformation helps improve cluster quality and that the conjunction of transformations leads to the best results. We also demonstrate that combining multiple clustering runs can greatly boost sound class discriminability.

2. Dirichlet process Gaussian mixture model

DPGMMs (also known as infinite GMMs) extend finite mixture models by the aspect of automatic model selection: The model

http://dx.doi.org/10.21437/Interspeech.2016-988

finds its complexity automatically given the data. Inference is typically sample based using a Markov chain Monte Carlo (MCMC) scheme such as Gibbs sampling. The sampler used here combines a restricted Gibbs sampler with a split/merge sampler. For more in-depth informations, please refer to [9, 18].

2.1. Generative process

Let $X = \{x_1, \cdots, x_n\}$ be a set of observations. The generative process of X given a DPGMM is as follows:

- Mixing weights $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ are generated according to a stick-breaking process
- GMM parameters $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_k\}$ are generated according to a prior distribution $\mathrm{NIW}(m_k, S_k, \kappa_k, \nu_k)$
- A label z_i is assigned to every x_i, according to $\boldsymbol{\pi}$
- x_i is generated according to the z_i-th GMM component

$\theta_k = \{\mu_k, \Sigma_k\}$ are Gaussian parameters, and the parameter set of the prior Normal-inverse-Wishart (NIW) distribution consists of a prior m_0 for μ_k, a prior S_0 for Σ_k, the belief-strength κ_0 in m_0 and the belief-strength ν_0 in S_0.

2.2. Inference

The parallelizable sampler alternates between a non-ergodic restricted Gibbs sampler and a split/merge sampler to form an ergodic MCMC sampler.

Restricted Gibbs sampling allows labels z_i to be sampled from a finite set Z. By definition of the DPGMM, the distribution of the mixture weights follows a Dirichlet distribution.

Split/merge sampling performs on the existing components. To provide good split candidates, each component is augmented with two sub-clusters c_{k_l} and c_{k_r}, and each observation of a component is augmented with a sub-cluster label $z_i^{sub} \in l, r$. Split moves are proposed in a Metropolis-Hastings fashion. Merge steps are proposed randomly.

2.3. Posteriorgram generation

The posterior probability of cluster c_k, given observation x_i is

$$p(c_k|x_i) = \frac{\pi_k N(x|\theta_k)}{\sum N(x|\theta_j)} \quad (1)$$

and $p_i = (p(c_1|x_i), \cdots, p(c_K|x_i))$ is the posteriorgram for x_i.

3. Unsupervised speech feature transformation

We showed in [15] that the quality of DPGMM based speech feature vector clustering can be improved by using LDA transformed features as input, where the LDA transformations were estimated in an unsupervised fashion. To further improve the clustering quality we propose an extension to this work by utilizing transformations that can be used in conjunction to benefit from additive effects. The transformations help to project feature vectors into a more suitable sub-space for sound class discrimination by feature de-correlation and speaker adaptation.

The need of labels and models makes it difficult to use these methods in a zero resource setting out of the box. Not only are there no labels for the target data available, but class identities and even the amount of classes are also unknown. Confronted with an unknown language there is often no easy way to bootstrap acoustic models. In order to overcome these issues, we use a two-staged clustering framework that automatically finds frame-based class labels in a first clustering of the target data.

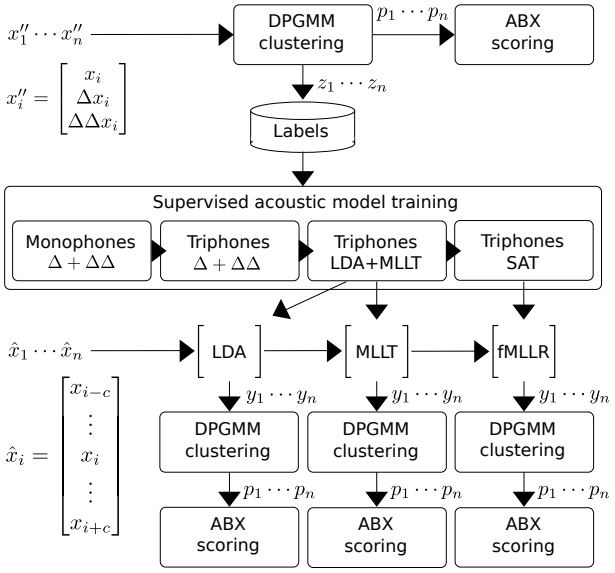

Figure 1: *Scheme of the iterative sampling process.*

3.1. Two-stage clustering

An initial DPGMM clustering on standard feature vectors with derivatives (x_i'') provides generic class labels and the hypothesized class membership of every speech frame. Each class is simply named with the numeric ID of the Gaussian that most likely produced the respective feature vector.

The frame-wise labels serve as basis for the subsequent model training. For each utterance, we collapse the labels to approximate a more natural textual reference type by compressing all subsequent tokens of the same type to a single token, imitating a phone based transcription of the audio recordings.

Figure 1 is a graphical overview of the iterative acoustic model training and clustering pipeline. Once we extract new feature vectors y_i using one or more of the transformations in conjunction, we perform another run of frame-based DPGMM clustering. Stages at which we extract the posteriorgrams for evaluation are named *ABX scoring*. Each stage produces a more advanced set of features for clustering, ranging from untransformed features to LDA+MLLR+fMLLR transformed features.

3.2. Supervised acoustic model training

We initialize the acoustic model by context-independent monophone training. Then we subsequently train context dependent triphones on untransformed standard features, followed by LDA and MLLT estimation and model training on transformed features. Finally, we train a SAT model with fMLLR.

We use a 3-state HMM topology with a skip from the first state to the next HMM to allow a more dynamic alignment. Due to the nature of the automatic labels, some utterances might be represented with a relatively high number of tokens. The skip state guarantees that an alignment is always found.

3.2.1. Dimension-reducing LDA

LDA is a simple linear transformation that we use to minimize intra-class discriminability and maximize inter-class discriminability of the speech features. LDA also enables us to do dimensional reduction of high-dimensional stacked feature vectors ($\hat{x}_i$) that span a larger context c by omitting lower-ranked

coefficients. Estimation of the transformation requires the feature vectors and respective class labels. In our pipeline we learn LDA transformations using the acoustic states as classes.

3.2.2. De-correlating MLLT

We attempt to apply MLLT to feature vectors so that correlations between the feature vector components are captured. MLLT is computed for distributions of speech observations in the HMMs of speech recognizers. The state-dependent transformations are estimated so that the likelihood of the adaptation data is maximized. We learn the transformations given the initialized HMMs of our unsupervisedly trained acoustic model.

3.2.3. Speaker-adapting fMLLR

fMLLR is an algorithm for speaker adaptive training (SAT). The idea of SAT is to capture inter-speaker variability in speaker dependent transforms and to generate speaker independent state distributions instead. The transformations are estimated based on alignments with speaker-independent features so that the likelihoods are maximized. We apply fMLLR in this zero resource setting because we expect the transformations to help eliminate variance caused by multiple speakers.

4. Clustering combination

For further improvement of the DPGMM clustering quality we developed a method to combine the results of n clustering runs, which are sets of posteriorgrams. For each frame, we add together the n individual posteriorgrams and normalize the new vectors so that they form proper posteriorgrams again.

Because the amount of found classes differs for each clustering run, a mapping between any two sets of posteriorgrams is needed. Given n sets of posteriorgrams, we randomly pick one of these sets as *target set*, and consider all other sets as *source sets*. We use the numeric frame-wise class labels as transcriptions for our data. For each source/target pair we first align the transcriptions and count the co-occurrences of classes. Then we keep the single most probable "translation" for each class and map the posteriorgrams into the target space as shown by the example in Figure 2.

Algorithm 1 Combination of dynamically sized posteriorgrams

Require: Set $\mathcal{P} = \{P_1, \cdots, P_n\}$ of sets of posteriorgrams
Ensure: Combined posteriorgrams $\hat{P}$

```
p_tgt ← random set from P
l_tgt ← generate labels from posteriorgrams p_tgt
P̂ ← p_tgt
for all p_src ∈ P \ p_tgt do
    l_src ← generate labels from posteriorgrams p_src
    count symbol pair occurrences in align(l_src, l_tgt)
    m ← 1-best mapping for all symbols in unique(l_src)
    P̂ ← P̂ + map(p_src, p_tgt, m)
end for
P̂ ← normalize P̂
```

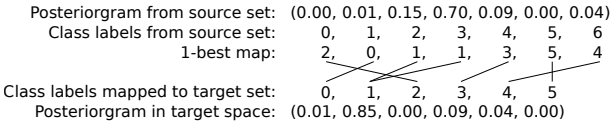

Figure 2: *Example of a posteriorgram mapping.*

5. Experiments

5.1. Data

The database for all our experiments is the official data set of the Interspeech zero resource speech challenge [1], which contains two separate data sets of pure speech for American English (4h 59min) and Xitsonga (2h 29min), a southern African Bantu language. The segments contain non-overlapping speech of exactly one speaker and noise or pauses. The English data is extracted from the Buckeye corpus and consists of conversational speech. The Xitsonga data is an excerpt of the NCHLT corpus and is comprised of read speech.

5.2. Evaluation

The evaluation metric we use to measure the cluster quality is based on the minimal pair ABX phone discriminability task [19], which is related to the ABX task used in psychophysics [20]. Each cluster is considered a phone in the context of the evaluation. We score GMM posteriorgrams that are computed for each speech frame after clustering, as described in Section 2.3. Let A and B be stimuli belonging to sound categories a and b. The ABX phone discrimination accuracy is

$$c(a,b) = \frac{1}{|a| \cdot |b| \cdot (|a|-1)} \sum_{A \in a} \sum_{B \in b} \sum_{X \in a \setminus \{A\}} \left(\delta_{d(A,X)<d(B,X)} + \frac{1}{2}\delta_{d(A,X)=d(B,X)}\right) \quad (2)$$

where $d(a,b)$ is the dynamic time warping (DTW) divergence and δ is an indicator function. As in Schatz et al. [19], we use the Kullback-Leibler divergence to compute the DTW divergences. Our scores are the error rates within and across speakers. The rates are averaged over all contexts for a given pair of central phonemes and then over all pairs of central phonemes.

5.3. Setup

We utilize the Kaldi speech recognition toolkit [21] to train the acoustic model used in our framework by following a standard scheme for speaker adaptive training.

We use the same parameters than Chen et al. [9] to ensure comparability. DPGMM sampling is done for 1500 iterations, and the priors are set so that m_0 is the global mean, S_0 is the global covariance, $\kappa_0 = 1$, and $\alpha = 1$. The value of ν_0 slightly varies and is set to the toolkit's default of $\nu_0 = D + 3$, where D is the dimension of the input feature vectors. All feature vector types are extracted for a frame length of 25msec and frame shift of 10msec. Mean variance normalization (MVN) and vocal tract length normalization (VTLN) is applied.

5.4. Baseline

For our baseline we extract 39 dimensional MFCC+Δ+$\Delta\Delta$ as input to the DPGMM sampler. We also compare to the results of Chen et al. [9] as reference due to the identical clustering setup. The details are listed in Table 1. Despite using the same sampling setup and input feature types, there is a mismatch between the results of Chen et al. [9] and our baselines. We believe this mismatch is caused by the fact that Chen et al. uses a custom voice activity detection for segmenting the full 10 hours of English data and does not mention any segmentation attempts for the 5 hours of Xitsonga data, where we use the officially provided segmentation that limits both data sets to about half the amount. Due to the differences we start with a higher error rate on English, but a lower error rate on Xitsonga.

	English		Xitsonga	
Features	within	across	within	across
MFCC+Δ+ΔΔ ([9])	10.8	16.3	9.6	17.2
MFCC+Δ+ΔΔ ([15])	12.2	19.5	8.9	14.2
PLP+Δ+ΔΔ	11.8	19.6	8.5	13.9
PLP+LDA	10.5	16.1	8.3	12.8
PLP+MLLT	10.5	16.2	8.4	12.9
PLP+MLLT+fMLLR	10.6	15.7	8.4	12.2
Best combination	10.0	14.9	8.1	11.7

Table 1: *The optimal results for each input feature type.*

5.5. Untransformed features

In our previous work [15] we found that PLP feature vectors are consistently leading to a higher clustering quality than MFCC feature vectors. We therefore conducted all clustering experiments based on this feature type.

5.6. Dimension-reducing LDA

The LDA transformation takes stacked standard feature vectors without their derivatives as input. Following our findings in [15] we fix the stacking context parameter set to $c = 4$, and the output dimensionality to $d = 20$. With the application of LDA we were able to produce feature vectors that considerably helped the DPGMM clustering process to find better clusters. The error rates for both languages dropped consistently, and especially across speakers a clear performance boost is observable. LDA features outperform our own baseline and also undercut the numbers of Chen et al. [9], thus compensating for the deficit in the baseline numbers that we had to begin with.

5.7. De-correlating MLLT

Applying MLLT to the LDA transformed features did not lead to a better clustering quality. This lets us assume that the decorrelating effects might not be able to aid this particular task, albeit being useful during a decoding task, as experience shows.

5.8. Speaker-adapting fMLLR

The transformations learned with fMLLR during the speaker adaptive acoustic model training prove to be very useful for boosting the discrimination capabilities across speakers. A relative improvement of 3% for English and almost 6% for Tsonga proves that fMLLR based speaker adaptive transformations can considerably improve clustering quality in the face of speaker variations and greatly benefit the clustering task.

5.9. Posteriorgram combination

We tested combining several clustering results across (a) transformations, (b) features, (c) LDA input and (d) output dimensionalities, and (e) given multiple identical clustering runs. The results are plotted in Figure 3.

For the combination across transformations, we ran separate clusterings for each type of transformed features. However, this seems not particularly helpful. Discrimination errors for English slightly drop, but increase for Xitsonga. As fMLLR is applied to LDA+MLLT features, all useful information seems already encapsulated in the fMLLR-transformed features. Similarly, combining separate MFCC and PLP feature clustering results led to lower errors on English, but not on Xitsonga.

Combining across LDA input dimensions, where the context parameter c ranges from 1 to 8, was by far the most efficient scheme, boosting the discrimination quality across languages considerably. Combination across LDA output dimensions $d \in \{16, 20, 23, 26\}$ also led to a better performance across the board, but to a lesser extent. These results let us assume that higher dimensional coefficients still carry complimentary information that can help in a combination scheme, albeit performing worse for clustering in isolation [15].

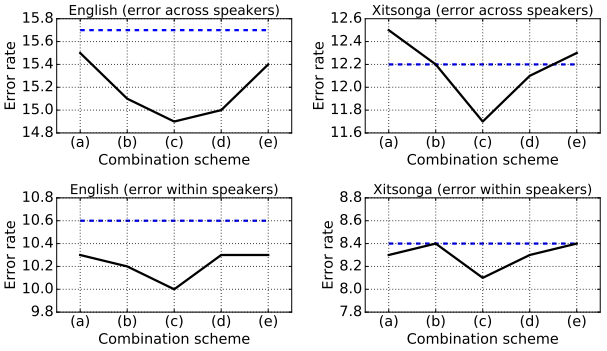

Figure 3: *Error rates after combination of clustering results. The dotted line is the best performance before combination.*

Clustering one feature set five times with subsequent combination of the results had a slightly positive effect on English, but was not helpful on Xitsonga. We assume the DPGMM cluster generally leads to consistent output and multiple parallel iterations are therefore not particularly useful.

6. Conclusion

We successfully utilized a supervised acoustic model training pipeline without supervision to improve DPGMM based feature vector clustering. Feature transformations estimated during model training can be used to map speech feature vectors into subspaces that are more suitable for clustering. Gaussian posteriorgrams extracted from a DPGMM that was sampled on transformed vectors carry better sound class discriminating characteristics than the ones sampled on untransformed standard features. We showed that LDA greatly benefits sound class discriminability within and across speakers. Consecutive fMLLR transformation noticeably decreases the discrimination error across speakers, proving the importance of speaker adaptation in solving the clustering task.

Combining multiple runs of LDA+MLLT+fMLLR-transformed feature vector clustering that use varying input dimensionalities yielded the the best results. We achieved error rates of 10% within and 14.9% across speakers for English, and 8.1% within and 11.7% across speakers for Xitsonga, respectively. Given these results, our proposed two-stage clustering framework clearly outperforms our own baseline, as well as the baseline set by Chen et al. [9].

Our two-staged clustering approach is particularly suitable for low-resource languages and the zero-resource scenario. Moreover, this framework might as well be of help for more general purposes beyond low-resource languages. In future work we will explore the applicability of our model training and clustering pipeline to solving other tasks beyond sound unit detection.

7. Acknowledgements

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and 26870371.

8. References

[1] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proceedings of Interspeech*, 2015.

[2] A. Park and J. Glass, "Towards unsupervised pattern discovery in speech," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 53–58.

[3] ——, "Unsupervised pattern discovery in speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 186–197, 2008.

[4] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008, pp. 165–168.

[5] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[6] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Association for Computational Linguistics Annual Meeting*, vol. 45, no. 1. Citeseer, 2007, p. 504.

[7] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 460–470.

[8] H. Kamper, A. Jansen, S. King, and S. Goldwater, "Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 100–105.

[9] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proceedings of Interspeech*, 2015.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[11] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 661–664.

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

[13] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 1137–1140.

[14] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[15] M. Heck, S. Sakti, and S. Nakamura, "Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario," in *Proceedings of SLTU*, 2016.

[16] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and k-means clustering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 521–528.

[17] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection." in *SDM*. SIAM, 2014, pp. 938–946.

[18] J. Chang and J. W. Fisher III, "Parallel sampling of dp mixture models using sub-cluster splits," in *Advances in Neural Information Processing Systems*, 2013, pp. 620–628.

[19] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proceedings of Interspeech*, 2013.

[20] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Psychology press, 2004, ch. 9.

[21] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi speech recognition toolkit," in *Proceedings of IEEE*, 2011.