Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline

Thomas Schatz[1,2], *Vijayaditya Peddinti*[3], *Francis Bach*[2],
Aren Jansen[3], *Hynek Hermansky*[3], *Emmanuel Dupoux*[1]

[1] LSCP, ENS/EHESS/CNRS, Paris, France
[2] SIERRA Project-Team, INRIA/ENS/CNRS, Paris, France
[3] HLT Center of Excellence, John Hopkins University, Baltimore, Maryland

thomas.schatz@ens.fr, vijay.p@jhu.edu, francis.bach@ens.fr, aren@jhu.edu,
hynek@jhu.edu, emmanuel.dupoux@gmail.com

Abstract

We present a new framework for the evaluation of speech representations in zero-resource settings, that extends and complements previous work by Carlin, Jansen and Hermansky [1]. In particular, we replace their Same/Different discrimination task by several Minimal-Pair ABX (MP-ABX) tasks. We explain the analytical advantages of this new framework and apply it to decompose the standard signal processing pipelines for computing PLP and MFC coefficients. This method enables us to confirm and quantify a variety of well-known and not-so-well-known results in a single framework.

Index Terms: zero-resource, speech representations, evaluation framework, minimal-pair ABX task

1. Introduction

Speech recognition technology crucially rests on adequate speech features for encoding input data. Several such features have been proposed and studied (MFCCs, PLPs, etc), but they are often evaluated indirectly using complex tasks like phone classification or word identification. Such an evaluation technique suffers from several limitations. First, it requires a large enough annotated corpus in order to train the classifier/recognizer. Such a resource may not be available in all languages or dialects (the so-called “zero or limited resource” setting). Second, supervised classifiers may be too powerful and may compensate for potential defects of speech features (for instance noisy/unreliable channels). However, such defects are problematic in unsupervised learning techniques. Finally, the particular statistical assumptions of the classifier (linear, Gaussian, etc.) may not be suited for specific speech features (for instance sparse neural codes as in Hermansky [2]). It is therefore important to replace these complex evaluation schemes by simpler ones which tap more directly the properties of the speech features.

Here, we extend and complement the framework proposed by Carlin, Jansen and Hermansky [1] for the evaluation of speech features in zero resource settings. This framework uses a Same-Different word discrimination task that does not depend on phonetically labelled data, nor on training a classifier. It assumes a speech corpus segmented into words, and derives a word-by-word acoustic distance matrix computed by comparing every word with every other one using Dynamic Time Warping (DTW). Carlin et al. then compute an average precision score which is used to evaluate speech features (the higher average precision, the better the features).

We explore an extension of this technique through Minimal-Pair ABX tasks (MP-ABX tasks) tested on a phonetically balanced corpus [3]. This improves the interpretability of the Carlin et al evaluation results in three different ways. First, the Same/Different task requires the computation of a ROC curve in order to derive average precision. In contrast, the ABX task is a discrimination task used in psychophysics (see [4], chapter 9) which allows for the direct computation of an error rate or a d’ measure that are easier to interpret than average precision [1] and involve no assumption about ROC curves. Second, the Same/Different task compares *sets of words*, and as a result is influenced by the mix of similar versus distinct words or short versus long words in the corpus. The ABX task, in contrast, is computed on *word pairs*, and therefore enables to make linguistically precise comparisons, as in word *minimal pairs*, i.e. words differing by only one phoneme. Variants of the task enable to study phoneme discrimination across talkers and/or phonetic contexts, as well as talker discrimination across phonemes. Because it is more controlled and provides a parameter and model-free metric, the MP-ABX error rate also enables to compare performance across databases or across languages. Third, we compute bootstrap-based estimates of the variability of our performance measures, which allows us to derive confidence intervals for the error rates and tests of the significance of the difference between the error rates obtained with different representations.

We provide technical details about our evaluation framework in Section 2 and apply it to the analysis of a pipeline of signal processing operations involved in the computation of the standard PLP [5] and MFC [6] coefficients in Section 3.

2. Methods

2.1. Stimuli

We used the CV subset of the Articulation Index Corpus (LDC-2005S22) [3], consisting in all possible Consonant-Vowel syllables of American English pronounced in isolation by 12 male and 8 female speakers, i.e., a total of 6839 stimuli recorded and sampled at 16KHz. We removed the silence surrounding each syllable through manual correction of the output of a speech activity detector.

2.2. Tasks

ABX tasks consist in presenting three stimuli A, B and X. A and B differ by some minimal contrast, and X is matched to either A or B. We use three variants of the task: in the *Phoneme across Talker* task (PaT), A and B differ by one phoneme (either the vowel or the consonant) and are spoken by the same talker. X is spoken by a different talker but has the same phonemes as either A or B. It measures talker invariance in phoneme discrimination. In the *Phoneme across Context* task (PaC), A and B differ only by one phoneme and are spoken by the same talker. X, also spoken by the same talker, matches A or B in one phoneme and differs from both in the other phoneme, measuring context invariance in phoneme discrimination. In the *Talker across Phoneme* task (TaP), A and B are spoken by two different speakers and are phonemically identical. X is spoken by the same speaker as either A or B, but differs from them by one segment, enabling the measurement of talker discrimination (see Table 1 for sample stimuli).

Table 1: *Example of a possible choice of the A, B and X stimuli for each MP-ABX task.* sp *stands for speaker.*

Task	A	B	X	Answer
PaT	/ba/ sp1	/ga/ sp1	/ba/ sp2	A
PaC	/ba/ sp1	/ga/ sp1	/gu/ sp1	B
TaP	/ba/ sp1	/ba/ sp2	/ga/ sp1	A

2.3. Model of the MP-ABX tasks

To perform these tasks on the basis of the speech representations a, b and x of the stimuli A, B and X, we begin as in [1] by computing the DTW distances $d(a, x)$ and $d(b, x)$ between A, X and B,X on the basis of an underlying frame-based distance metrics. Then, the sign of $d(a, x)-d(b, x)$, is used to determine the response of the model (respectively B or A for a positive or negative sign) and an error rate is computed. The choice of the underlying frame-based metrics is important and may impact the results. Here, we follow the recommendation of [1] and use the cosine distance in all our tests.

2.4. Analyses

The error rate score for a given MP-ABX task is defined as the average error rate over all the relevant triplets of stimuli A, B and X in the database. For the PaT and PaC tasks, we additionally compute average error rates over consonantal or vocalic constrats. We compute confidence intervals for these average error rates by resampling across talkers. We also resample across talkers to perform significance tests when we test error rates differences.

2.5. The classical MFC/PLP signal processing pipeline

As in [7], we apply our evaluation framework to representations obtained at various stages of a speech processing pipeline leading to standard MFC [6] or PLP [5] coefficients with or without RASTA filtering [8]. We start from a short-term power spectrum representation of the speech waveform obtained through a Fast Fourier Transform of 25ms frames taken each 10ms. Then we form one of 16 representations by making a succession of 4 binary choices (see Figure 1): use a linear or a Mel frequency scale; weight frequency channels according to human's equal-loudness contour or not; cubic root compress the dynamic range of frequency channels or not; apply RASTA filtering or not. We study these representations with 2, 5, 8, 13, 22, 36, 60 and 100 frequency channels. To complete our study of the MFC/PLP pipeline we apply Linear Predictive Coding to some of these 16 representations and re-estimate a cepstrum from the filter coefficients. In particular, we obtain standard PLP coefficients through the following path in the pipeline: *Mel/equal-loudness/compression/no RASTA/LPC/cepstrum estimation*. We also apply a cepstral transform (log plus DCT transform) to some of the representations and obtain the standard MFC coefficients through the following path in the pipeline: *Mel/no equal-loudness/no compression/no RASTA/cepstral transform*. We study these representations with 2, 5, 8, 13, 22, 60 and 100 cepstral coefficients. These pipelines were adapted from Dan Ellis' audio toolbox [9].

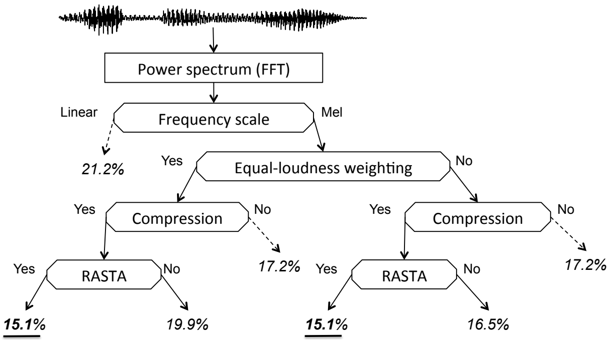

Figure 1: *First stages in processing pipelines for the computation of standard MFC and PLP coefficients. The MP-ABX error rate for the PaT minimal pair discrimination task is in italics. The best pipelines are shown with plain arrows, and the best scores are underlined. Parts of the pipeline not shown are indicated by dashed arrows and the best error rate achieved in each hidden part is indicated next to the arrow.*

3. Results

3.1. First stages of the MFC/PLP pipeline

We first analyze the results for the 16 spectral representations represented on Figure 1. We begin with the effect of the number of spectral channels on the MP-ABX error rate (Figure 2). For a simple Mel-spectrum (Figure 2 (a)), the optimal number of channels is highest in the TaP task (36), intermediate in the PaC task (13) and lowest in the PaT task (8). The difference between the error rate for the optimal number of channels and the error rates for neighboring number of channels is small in all three tasks, but we find that it is significant for the PaC and PaT tasks when resampling across talkers. This means that these precise optimal values can be found robustly across talkers. We also observed that the optimal number of spectral channels is consistently higher in the PaC task than in the PaT task for the 8 representations from Figure 1 that are derived from a Mel-spectra (Figure 2 (b)). These results are coherent with previous findings [5] that speaker-specific information is contained in the fine details of the spectra, so that coarser spectral resolution yields features more invariant to speaker change.

Next, we compare the error rates of the different representations in the PaT task (Figure 1). The number of spectral channels is optimized for each feature independently. Using a Mel-scale is clearly beneficial: the worst error rate for a represen-

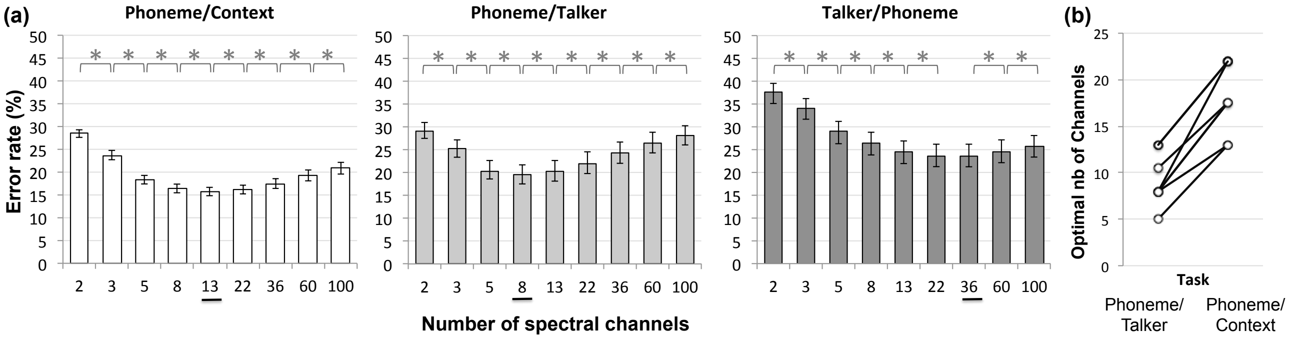

Figure 2: *(a) Average MP-ABX error rate in each task for a simple Mel spectrum with various number of spectral channels. Error bars represent 95% confidence intervals (sampled across talkers). The optimal number of channels is underlined and differences between error rates for adjacent number of channels that are significant at a level $\alpha = 1\%$ are indicated by a star. (b) Optimal numbers of channels in the PaC and PaT tasks for the* 8 *representations from figure 1 that are derived from a Mel-scale.*

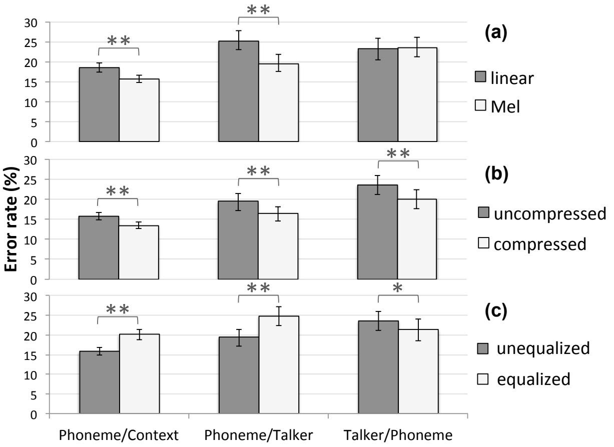

Figure 3: *MP-ABX error rate in the three tasks for (a) a simple Mel-scale or linear scale spectrum, (b) a Mel-scale spectrum with or without cubic root compression, (c) a Mel-scale spectrum with or without equal-loudness weighting. Error bars represent 95% confidence intervals (sampled across talkers). Differences significant at a level $\alpha = 5\%$ and $alpha = 1\%$ are indicated by one and two stars respectively.*

tation using a Mel-scale (19.9%) is better than the best error rate for a representation using a linear scale (21.2%). The best representations are also consistently obtained when using cubic root compression and RASTA filtering, which yield improvements of 2.1% and 1.4% respectively of the error rate for the best representation. The other effect we observe is more surprising: equal-loudness filtering has a detrimental effect (3.4% increase in error rate) in the absence of RASTA filtering. When RASTA filtering is applied the effect of equal-loudness filtering is wiped out.

We now look at the effect of the frequency scale, cubic-root compression of the dynamic range and equaI-loudness weighting in each task (Figure 3). Using a Mel-scale benefits strongly to phoneme discriminability both across talker and contexts and does not affect the ability to discriminate speakers. Cubic-root compression benefits to phoneme discriminability across talker and contexts too, but also to speaker discriminability. The adverse effect of equal-loudness filtering on phoneme discriminability occurs also across contexts and coincides with a slight improvement of talker discriminability.

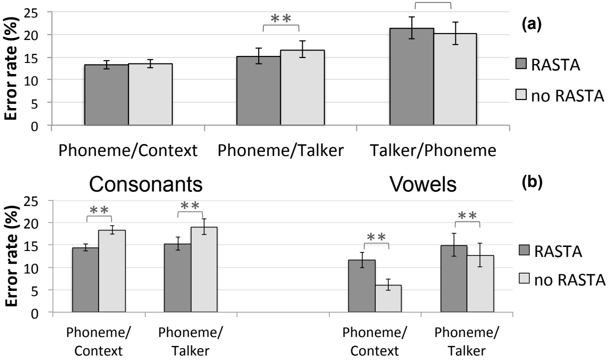

Figure 4: *(a) MP-ABX error rate in the three tasks for a cubic-root compressed Mel-scale spectrum with RASTA filtering or not. (b) Consonantal and vocalic MP-ABX error rates in the PaC and PaT tasks for the same representations. Error bars represent 95% confidence intervals (sampled across talkers). Differences significant at a level $\alpha = 5\%$ and $alpha = 1\%$ are indicated by one and two stars respectively.*

We next study the effect of RASTA filtering in the three tasks (Figure 4(a)). We start from our best representation so far: a cubic-root compressed Mel-sepctrum. RASTA filtering improves the discriminability of phoneme across talkers at the same time that it impairs discriminability of talkers across phonetic contexts thus performing a form of speaker normalization. This is also supported by the absence of significant difference between RASTA filtering and mean-variance normalization on the discriminability of phonemes across contexts. We uncover additional details on the coding properties of RASTA filtering by looking at error rates for consonants and vowels separately (Figure 4 (b)). RASTA filtering improves consonant coding and impairs vowels coding across both contexts and talkers. Moreover, while RASTA filtering improves consonant coding in both tasks by a comparable amount (3.7% and 3.4%) it impairs vowel coding by a lesser amount in the PaT task (1.8%) than in the PaC task (4.5%). All these results are coherent with the view of RASTA filtering as a form of short-term adaptation, enhancing transients in the signal that are useful for discriminating consonants and removing speaker-specific steady-state information, which is helpful in discriminating vowels within a given talker but less so across talkers.

3.2. Standard MFC and PLP coefficients

We now investigate the final steps of the pipeline. First we study the effect of the number of cepstral coefficients for standard MFCC (Figure 5 (a)). The number of spectral channels had a very small effect on the error rate in the three tasks with a range of variation lower than 0.9% in the PaC, 1.2% in the PaT task and 3% in the TaP task for any given number of cepstral coefficients. By contrast, the number of cepstral channel has a much bigger effect (Figure 5 (b)), similar to that of the number of spectral channels in the absence of a cepstral transform (Figure 2 (a)). Best results were obtained with 22 spectral channels in the Phoneme/Talker task, 60 in the Phoneme/Context task and 100 in the Talker/Phoneme task with respectively 13, 8 and 36 cepstral coefficients, coherent with the idea that a coarser spectral and cepstral resolution increases talker invariance and in striking accord with usual choices for these parameters.

Next, we compare standard MFC and PLP features error rates on the three tasks. In these and subsequent results each representation is computed with 47 spectral channels (1 channel per Mel) and 13 cepstral coefficients. MFCC (Table 2, 1) outperform PLP coefficients (Table 2, 2) on all tasks. To test whether this is due to the detrimental effect of equal-loudness filtering previously found we tested PLP coefficients computed without equal-loudness filtering (Table 2, 3). Now PLP coefficients are slightly better than MFCC except on the TaP task. We next look at MFCC computed with cubic-root compression (Table 2, 4) and PLP computed without it (Table 2, 5). There is no clear pattern of improvement or worsening in the result. This may be because the logarithm of the spectra is taken to obtain cepstral coefficients, which constitute a form of compression of the dynamic range, so that the benefits of doing an additional cubic-root compression are not clear. We also see that the benefit of using a Mel-scale (Table 2, 1) instead of a linear scale (Table 2, 6) and the talker normalization effects of using RASTA filtering (Table 2, 7) carry on to the cepstral domain.

Table 2: *MP-ABX error rates (%).*

	Feature	PaC	PaT	TaP
1	standard MFC	13.7	17.8	17.7
2	standard PLP	14.2	18.3	18.9
3	PLP unequalized	**13.6**	17.6	19.5
4	MFC compressed	13.7	18.1	18.2
5	PLP uncompressed	14.2	17.6	20.6
6	MFC linear	17	24.9	**16.2**
7	MFC RASTA	13.8	**16.7**	21

4. Conclusion

We built upon previous work by Carlin, Jansen & Hermansky [1] to propose a new framework for the evaluation of speech representations in the zero or low resource setting. We used several MP-ABX tasks to provide rich and easily interpretable information about the coding properties of each representation. We demonstrated the effectiveness of our framework by applying it to a pipeline of signal processing operations involved in the computation of standard MFC and PLP coefficients. We were able to confirm quantitatively some-well known results, such as the talker-normalization properties of RASTA filtering and also uncover a few unexpected results, such as the detrimental effect of equal-loudness weighting of the frequency channels on the discriminability of phonemes. In future work the coding properties of speech representations could be further characterized by looking at more detailed aspects, such as, for example, the coding of specific linguistic features. Also, other classical signal processing techniques (e.g. mean-variance normalization, cepstral liftering or delta-coefficients), more sophisticated models (e.g. [10, 11, 12, 13, 14, 15, 16, 17, 18]), other metrics [19] as well as human performance could be investigated within our framework.

5. Acknowledgements

We thank Anna Kolesnikov, Tomas Bergvelt, Sharon Peperkamp, Anne Christophe and Daniel Pressnitzer. The research leading to these results received funding from the European Research Council under the FP/2007-2013 program / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-11-0001-02 PSL*, ANR-10-LABX-0087) and from the Fondation de France.

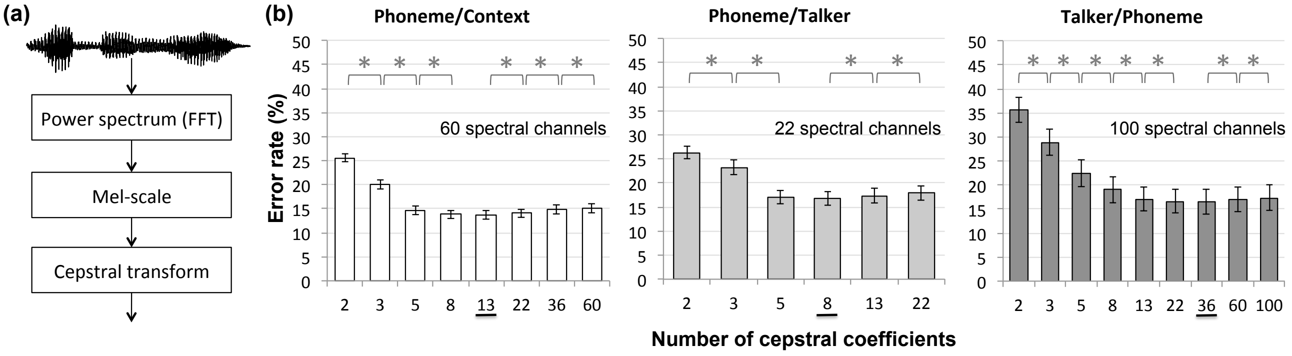

Figure 5: *(a) Processing steps for computing standard MFCC in our pipeline. (b) MP-ABX error rate in the three tasks for various numbers of cepstral coefficients, for classical MFCC. Error bars represent 95% confidence intervals (sampled across talkers). The optimal number of coefficients is underlined and differences between error rates for adjacent number of coefficients that are significant at a level $\alpha = 1\%$ are indicated by a star. The number of spectral channels was chosen for each task to optimize the minimal error rate in that task.*

6. References

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

[2] G. Sivaram and H. Hermansky, "Sparse multilayer perceptron for phoneme recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 23–29, 2012.

[3] P. Fousek, P. Svojanovsky, F. Grezl, and H. Hermansky, "New nonsense syllables database analyses and preliminary asr experiments," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2004, pp. 2004–29.

[4] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Lawrence Erlbaum, 2004.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[6] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 91–103, 1976.

[7] J. Rajnoha and P. Pollák, "ASR systems in noisy environment: Analysis and solutions for increasing noise robustness," *Radioengineering*, vol. 20, no. 1, pp. 74–84, 2011.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

[10] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C.-y. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JH CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proceedings of ICASSP 2013*, 2013.

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

[12] 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*, 2008, pp. 165–168.

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

[14] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of ACL*, 2012.

[15] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in speech, hearing and language processing*, vol. 3, pp. 547–563, 1996.

[16] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'82.*, vol. 7, 1982, pp. 1282–1285.

[17] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," in *Readings in speech recognition*, 1990, pp. 101–111.

[18] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[19] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.