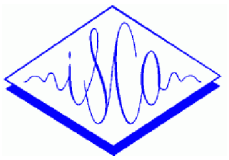

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Intrinsic Spectral Analysis for Zero and High Resource Speech Recognition

Aren Jansen, Samuel Thomas, Hynek Hermansky

Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, Maryland, USA
aren@jhu.edu, sthomas@jhu.edu, hynek@jhu.edu

Abstract

The constraints of the speech production apparatus imply that our vocalizations are approximately restricted to a low-dimensional manifold embedded in a high-dimensional space. Manifold learning algorithms provide a means to recover the approximate embedding from untranscribed data and enable use of the manifold's intrinsic distance metric to characterize acoustic similarity for downstream automatic speech applications. In this paper, we consider a previously unevaluated nonlinear out-of-sample extension for intrinsic spectral analysis (ISA), investigating its performance in both unsupervised and supervised tasks. In the zero resource regime, where the lack of transcribed resources forces us to rely solely on the phonetic salience of the acoustic features themselves, ISA provides substantial gains relative to canonical acoustic front-ends. When large amounts of transcribed speech for supervised acoustic model training are also available, we find that the data-driven intrinsic spectrogram matches the performance of and is complementary to these signal processing derived counterparts.

Index Terms: intrinsic spectral analysis, manifold learning, speech recognition, zero resource

10.21437/Interspeech.2012-266

1. Introduction

The dominant paradigm in the speech recognition community for the past four decades has been to train automatic systems with as much transcribed data we can get our hands on. This strategy has led to the development of highly accurate systems that have begun to find a place in our daily lives. An unfortunate consequence of this trajectory, however, is that state-of-the-art recognition performance can only be achieved on languages and domains for which vast transcribed training resources either exist or can be easily obtained. When forced to deal with resource impoverished settings, there is a large gap between expectations and capability when we apply state-of-the-art technologies.

The most common approach to accommodating low resource language scenarios has been the introduction of semi-supervised learning methods in the acoustic and language models. A less studied alternative strategy is to approach the problem in the front-end, where we can use untranscribed speech to learn feature transformations that can replace or supplement the traditional signal processing techniques behind mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP). Commonly used (linear) approaches in this vein include principal components analysis (PCA) and, given access to training data, heteroscedastic linear discriminant analysis (HLDA). This front-end strategy is most relevant in the zero resource setting, where we only have the untranscribed evaluation data itself and no access to transcribed speech, acoustic/language models, or pronunciation dictionaries. However, even in this maximally resource impoverished setting, there is no barrier to considering unsupervised data-driven front-ends such as intrinsic spectral analysis (ISA) [1].

ISA attempts to learn, from untranscribed speech alone, the underlying manifold structure that speech sound spectra (or other acoustic representations) are restricted to as a consequence of the physics of production [1, 2]. At a high level, the goal is to leverage the class-independent distribution of the data to learn an embedding that will be more faithful to the acoustic-phonetic structure of the language and factor out speaker and channel dependencies. Past studies [1, 3, 4] considered a variant of ISA that makes a linear approximation, which prevents learning a transformation that accurately models manifolds with high extrinsic curvature. In this paper, we evaluate ISA with nonlinear out-of-sample extension, both on a zero resource spoken term discovery task and high resource phonetic recognition task using a state-of-the-art acoustic model back-end. We begin with an overview of the technique.

2. Intrinsic Spectral Analysis

The Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ on a Riemannian manifold $\mathcal{M}$ is the (positive semi-definite) second order differential operator. Its eigenfunctions form an orthogonal basis for square integrable functions defined on the manifold, i.e., if $\{\lambda_i\}$ and $\{e_i\}$ are the sorted eigenvalues and corresponding eigenfunctions, respectively, then any function $f : \mathcal{M} \rightarrow \mathbb{R}$ may be written $f = \sum_i a_i e_i$ for some $\{a_i\}$. With the intuition that second derivatives measure the degree of function curvature, the Laplace-Beltrami operator may be used to quantify the smoothness for functions defined on the manifold. Given a measure μ on the manifold, the functional

$$S[f] = \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 d\mu = \langle \Delta_{\mathcal{M}} f, f \rangle_{\mathcal{L}^2(\mathcal{M})} \qquad (1)$$

increases as smoothness of f decreases [5]. Here, $\nabla_{\mathcal{M}}$ is the gradient operator on $\mathcal{M}$ and $\langle \cdot, \cdot \rangle_{\mathcal{L}^2(\mathcal{M})}$ is the $\mathcal{L}^2$ inner product on $\mathcal{M}$. It follows that the smoothness of an eigenfunction is determined by the magnitude of the corresponding eigenvalue, since $S[e_i] = \lambda_i$. Thus, initial eigenfunctions $e_i : \mathcal{M} \rightarrow \mathbb{R}$ vary most smoothly with *geodesic* distance on $\mathcal{M}$ and best preserve locality on the manifold regardless of the particular form of the embedding. In this way, they define an optimal embedding and a natural coordinate system for the manifold.

The discrete analogue to the continuous Laplace-Beltrami operator is the graph Laplacian, $\mathbf{L}$, a positive semi-definite matrix that satisfies (in matrix form) all the properties given above for $\Delta_{\mathcal{M}}$ [5]. Given a manifold $\mathcal{M}$ embedded in a Euclidean acoustic feature space $\mathbb{R}^d$ and a collection of n points $X = \{x_1, \ldots x_n\} \subset \mathcal{M}$, the graph Laplacian is defined as follows. First, we construct a weighted, undirected adjacency

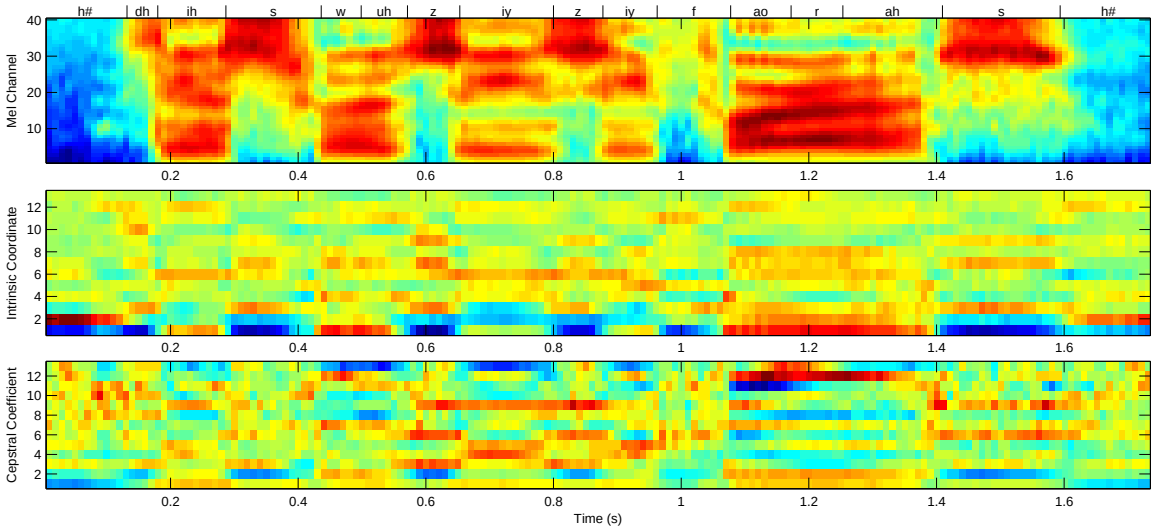

Figure 1: Log mel spectrogram (top), intrinsic spectrogram (middle), and MFCCs for the utterance "This was easy for us."

graph $G = (V, E)$ with one vertex $V_i \in V$ per data point $x_i \in X$, and connect vertices V_i and V_j with an edge of similarity weight W_{ij} if x_i is one of the κ nearest neighbors of x_j or x_j is one of the κ nearest neighbors of x_i. The (normalized) graph Laplacian is then defined as the $n \times n$ matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{D}$ is a diagonal matrix with elements $D_{ii} = \sum_j W_{ji}$ (the degree of vertex V_i). The basis determined by the graph Laplacian serves as an approximation to the basis for $\mathcal{M}$. However, unlike the Laplace-Beltrami eigenbasis, the graph Laplacian eigenvectors are defined only on the samples of X, not the entire manifold. Clearly, for speech front-end applications, we cannot compute new intrinsic projection maps for every novel utterance.

Intrinsic spectral analysis, as originally presented [1], addresses this problem with an out-of-sample extension using an interpolation scheme based on an unsupervised variant of the semi-supervised manifold regularization framework, presented in [6]. In the unsupervised learning setting, the algorithm input is a set of unlabeled acoustic feature vectors, X, that form a mesh of data points that lie on the manifold. The optimization problem takes the form

$$f^* = \arg \min_{f \in \mathcal{H}_K} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L} \mathbf{f}, \tag{2}$$

where $\mathcal{H}_K$ is the RKHS for some positive semi-definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\mathbf{L}$ is the graph Laplacian, and $\mathbf{f} = \langle f(x_1) f(x_2) \dots f(x_n) \rangle^T$ is the column vector of values of f computed on the graph. The first term is the extrinsic norm, limiting the complexity of the solution in the original ambient space. The second term is graph analogue of the intrinsic smoothness functional of Equation 1. The tunable parameter ξ determines the balance between extrinsic and intrinsic smoothness of the functions determined. By the RKHS representer theorem [6], the j-th component of our new projection map is

$$f_j^*(v) = \sum_{i=1}^{n} \alpha_i^{(j)} K(x_i, v), \tag{3}$$

where $\{x_i\}$ are the input unlabeled data, and $\alpha^{(j)} \in \mathbb{R}^n$ is the j-th eigenvector (sorted by eigenvalue) to the generalized eigenvalue problem

$$(\mathbf{I} + \xi \mathbf{L}\mathbf{K})\alpha = \lambda \mathbf{K} \alpha. \tag{4}$$

Here, $\mathbf{K}$ is the $n \times n$ Gram matrix defined on the input unlabeled data by $K_{ij} = K(x_i, x_j)$. Our set of projection maps is now defined out-of-sample, i.e., f_j^* may be evaluated at an arbitrary point in $\mathbb{R}^d$. To summarize, ISA is accomplished using the following three steps:

1. Given a set of unlabeled acoustic feature vectors $X = \{x_i\}_{i=1}^n$, compute the graph Laplacian $\mathbf{L}$.
2. Given a kernel K, solve the generalized eigenvalue problem of Equation 4 for the weights $\{\alpha^{(j)}\}_{j=1}^{d'}$.
3. Project novel feature vectors onto the first d' intrinsic basis functions according to Equation 3.

Solving the generalized eigenvalue problem for large n is computationally intensive, but need only be performed once offline using a random sample of frames. Subsequent intrinsic spectrogram generation for a novel utterance requires only the computation of Equation 3 and is a scalable linear time operation.

While the eigenfunctions of the graph Laplacian represent nonlinear projection maps, the functional complexity of the ISA out-of-sample extension will be limited by the choice of kernel function K. Previous studies of ISA were limited to a linear kernel of the form $K(x, y) = x^T y$. To accommodate nonlinear intrinsic projections maps for extrinsically curved manifolds, we perform the optimization of Equation 3 over the RKHS for the radial basis function (RBF) kernel, $K(x, y) = \exp\left(-\|x - y\|^2 / 2\sigma^2\right)$, capable of modeling highly complex functions.

Figure 1 displays a log mel spectrogram (40 channels) for a TIMIT utterance (sx3), along with the corresponding intrinsic spectrogram (13 components) and mel frequency cepstral coefficients (13 components). Much like the MFCCs, the intrinsic spectrogram compresses the distributed spectral information into a smaller number of dimensions. Much like principal component analysis, ISA produces features with decreasing component-wise variance, with most of the total variance concentrated in the first ten dimensions. Figure 2 displays the correlation of each intrinsic coordinate with each of the phonetic classes, computed over the entire TIMIT corpus. Both the raw and thresholded (at zero) correlations are shown. We find that many of the individual intrinsic coordinates can be understood according to some broad phonetic class distinction (or, alternatively, with distinctive feature theory). For example, the first

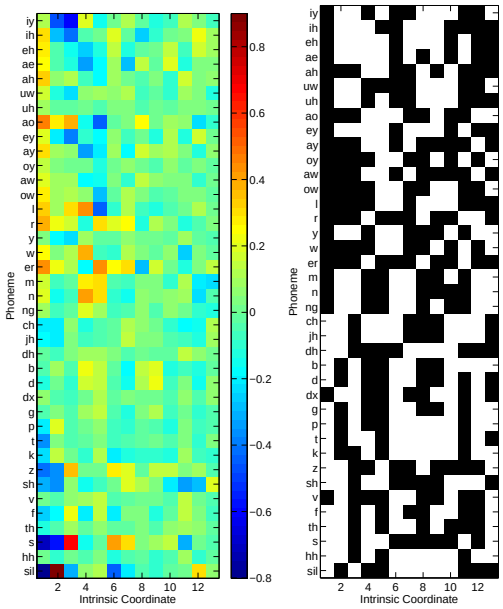

Figure 2: Correlation (left) and thresholded correlation (right) between phonetic classes and intrinsic spectral features.

intrinsic coordinate clearly separates sonorant from obstruent phonemes. Given obstruent class membership, the second intrinsic coordinate distinguishes stops and fricatives. Note that if we view each row of the thresholded correlation matrix as a binary vector, low pairwise Hamming distance indicates similar phoneme pairs (e.g. /jh/ and /ch/).

3. Experiments

We are interested in determining the utility of an ISA-based front-end for speech recognition tasks. Past studies [1, 3, 4] have considered only the linear out-of-sample variant on various toy applications. Here we use the extremely well studied TIMIT corpus that, while not representing a very difficult speech recognition challenge, provides a well-understood basis to evaluate *nonlinear* ISA against mainstream front ends. On one extreme, we consider the use of nonlinear ISA on a completely unsupervised task of spoken term discovery. On the other extreme, we consider a highly supervised phonetic recognition evaluation using multi-layer perceptron based acoustic models.

3.1. Unsupervised Term Discovery

Unsupervised spoken term discovery is the task of searching potentially large untranscribed speech collections for repeated words and phrases without using any language specific resources other than the collection itself. The prevailing approaches [7, 8, 9] rely on exhaustive dynamic time warping-based searches across all pairs of speech intervals drawn from the collection. While a computationally challenging endeavor, recent efficient approximation strategies [10] have made processing hundreds of hours of speech tractable. Moreover an evaluation metric has been defined [11] that circumvents the need to carry out the exhaustive search.

The main stumbling block for the task is speaker independence since the ability to associate acoustic realizations of a given phoneme or word spoken by different speaker is exactly what acoustic model supervision is meant to accomplish. In [11], it was found that signal processing-derived front-ends like PLP and MFCC were of comparable utility, both falling far short of supervised representations like matched-language posteriorgrams. The question for the present study is whether ISA recovers a representation that, when coupled with a suitable distance metric, will provide increased speaker independence while maintaining phoneme and word discriminability.

Table 1: Spoken term discovery evaluation results (in %). Distance metrics listed are those used in graph construction.

Features	Dimension	AP
Mel Spectrogram	40	10.7
Log Mel Spectrogram	40	7.6
Linear ISA-PLP (cosine)	39	28.2
Linear ISA-LogMelSpec (cosine)	39	25.6
PLP	39	34.8
MFCC	39	33.8
Nonlinear ISA-PLP (Euclidean)	39	40.0
Nonlinear ISA-PLP (cosine)	39	45.9
Nonlinear ISA-LogMelSpec (Euclidean)	39	38.6
Nonlinear ISA-LogMelSpec (cosine)	39	**48.5**
English Posteriorgram	40	75.4

We evaluated this question with the procedure defined in [11], adapted to the TIMIT corpus. Using the time aligned word transcriptions, we extracted all word examples that were at least 0.35 s in duration and at least 6 characters as text. Across the entire TIMIT corpus, this produced approximately 11k examples across 3,745 word types. We proceeded by computing all pairwise dynamic time warping (DTW) distances, involving some 60 million DTW computations (completes in approximately 10 minutes on a cluster of 100 CPUs). Finally, treating DTW distance as a same/different classifier score, we computed the average precision (AP) for the task of separating same word type pairs from different word type pairs. Due to its proven success for this task [11], we considered cosine distance as the DTW frame level distance metric for all features evaluated. Note that this choice removes any benefit of PCA.

While nonlinear ISA requires no labeled data for training, there are three parameters that must be chosen by the user: κ, the number of nearest neighbors for graph construction; ξ, the weight on the intrinsic norm in the optimization problem; and σ, the width of the RBF kernel (for nonlinear ISA only). Ideally, the dependence of downstream performance on the values of these parameters would be weak, precluding the need for labeled development data. In practice, we found $\kappa = 10$, $\xi = 30$, and $\sigma = 0.4m$, where m is the mean distance between samples in X, to generally work well. Increasing ξ beyond 30 led to a negligible loss in performance, indicating that the ambient regularizer is of minimal utility. Performance was moderately dependent on κ and σ; varying κ from 4 to 12 and σ from $0.1m$ to $1.0m$ resulted in at most 4% loss in average precision.

Table 1 lists this average precision (AP) for several front-ends. All features were computed in 25 ms windows sampled every 10 ms. The mel-scale spectrogram used 40 mel channels and the PLP used a 12th-order LPC-smoothed, bark-scale spectrogram. Thirteen cepstral coefficients (including DC component) were used for both PLP and MFCC. For the spectrograms, PLP, and MFCC, each feature dimension was normalized to zero mean and unit variance. Intrinsic spectrograms were derived from both the PLP (ISA-PLP) and the log mel spectrogram (ISA-LogMelSpec). In both cases, n =10k samples (chosen completely at random) were used to construct the graph Laplacian. We kept only the first 13 intrinsic components (skipping the first trivial dimension) to maintain the same feature dimension as PLP and MFCC. Velocity (Δ) and acceleration ($\Delta\Delta$) features were included for all of these features.

Table 2: Phone recognition results (all in % phone accuracy).

Features	Accuracy
MFCC	76.8
PLP	77.0
Nonlinear ISA-LogMelSpec	76.0
Nonlinear ISA-PLP	76.7
DS: PLP + Nonlinear ISA-LogMelSpec	**78.5**
DS: PLP + Nonlinear ISA-PLP	78.0

Finally, to set a supervised performance ceiling, we also considered English posteriorgram features generated using a multilayer perceptron with a single hidden layer of size 2000 nodes trained on all sx/si sentences (input PLP w/ 9-frame context).

There are several trends apparent in the results of the evaluation: (1) Not surprisingly, canonical front ends like PLP and MFCC are much better than the raw mel frequency spectrogram for cross speaker word matching, though both fall short of fully supervised phonetic posteriors by approximately 40% AP. (2) Linear ISA improves upon the log mel spectrogram, but falls short of PLP and MFCC. Nonlinear ISA is substantially better than the linear version, a consequence of the curvature of the speech manifold. (3) Regardless of whether we derive our nonlinear intrinsic spectrograms from PLP or mel spectrograms, we still end up with substantially improved representational speaker independence relative to both PLP and MFCC features. The improvement in cross-speaker word matching average precision is nearly 15% absolute (more than 40% relative), representing a third of the gap between the original PLP/MFCC and the supervised performance ceiling. (4) While Euclidean distance is the metric of choice for virtually all supervised speech recognition applications, it was clearly demonstrated in [11] that cosine distance provides a more natural notion of similarity in the zero resource regime. Likewise, we find substantially better performance when we construct the graph Laplacian using cosine distance instead of Euclidean distance.

3.2. Supervised Phonetic Recognition

Next, we considered the more familiar task of supervised phonetic recognition on the TIMIT corpus, employing a state-of-the-art hidden Markov model/multi-layer perceptron (HMM/MLP) back-end to evaluate ISA against traditional front-ends. First, we used each feature type to train a hierarchical MLP system [12] which estimates posterior probabilities of phoneme classes. The hierarchical MLP was trained using two levels of neural networks. At the first level, tri-state phoneme posterior probabilities of 49 speech sounds were estimated using 3-layer MLPs on the each feature set with a 9-frame context. The second MLP level in the hierarchy was trained on the tri-state posterior probabilities outputted by the first MLP with a large context of 23 frames. The estimated posterior probabilities were then converted to scaled likelihoods and used along with 3-state HMMs with equal self and state transition probabilities to model each phoneme class. The Viterbi algorithm is finally used along with a bigram language model to decode test phoneme sequences using these hybrid models. Phoneme recognition accuracies are reported on the standard 39 phoneme classes by the standard reduction from the 49-class decoding set. We also considered the combination of multiple feature types, which we performed on the final output posteriorgram of each constituent stream using the Dempster-Shafer (DS) theory of evidence [13]. Further details can be found in [12].

Table 2 lists the phone recognition accuracies using the same front-ends considered in Section 3.1 coupled to the above-described back-end. For ISA, we determined suitable parameter values using cross-validation ($\kappa = 6$, $\xi = 30$, $\sigma = 1.0m$). We find that the nonlinear ISA front-ends match the performance of the traditional MFCC and PLP front-ends and display substantial complementarity when DS combined. Without supervision of any kind, intrinsic spectral analysis is able to recover a nonlinear transformation that, when applied to the mel spectrogram, produces features as useful for subsequent recognition as linear prediction and cepstral analysis. This is a remarkable result, since once we construct the nearest neighbor graph with binary weights, all absolute notions of locality have been obscured to the optimization. Thus the topological structure alone of our unlabeled data sample is sufficient to recover a representation that encodes the necessary information for successful recognition downstream. The hope is that as one moves to new languages and non-speech applications, this sort of data driven approach will work as-is, without having to redesign the signal processing algorithms for each new task.

4. Conclusions

We have conducted a thorough study of intrinsic spectral analysis with nonlinear out-of-sample extension for downstream speech recognition applications. We have demonstrated unprecedented zero resource word matching performance and state-of-the-art high-resource phonetic recognition performance. Moreover, once the projection maps are computed in an offline optimization, nonlinear ISA is a linear time algorithm and is thus amenable to large scale applications.

5. References

[1] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds," in *Proceedings of ICASSP*, 2006.

[2] R. Togneri, M. D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," in *Proc. of IEE Conf. on Communications, Speech and Vision*, 1992.

[3] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *Proc. of ICASSP*, 2008.

[4] F. Tompkins and P. J. Wolfe, "Approximate intrinsic Fourier analysis of speech," in *Proc. of Interspeech*, 2009.

[5] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Represenation," *Neural Computation*, vol. 16, pp. 1373–1396, 2003.

[6] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

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

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

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

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

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

[12] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. Interspeech*, 2007.

[13] F. Valente and H. Hermansky, "Combination of acoustic classifiers based on Dempster-Shafer theory of evidence," in *Proc. ICASSP*, 2007.