Representing Nonspeech Audio Signals through Speech Classification Models

Huy Phan[⋆†], *Lars Hertel*[⋆], *Marco Maass*[⋆], *Radoslaw Mazur*[⋆], and *Alfred Mertins*[⋆]

[⋆]Institute for Signal Processing, University of Lübeck, Germany
[†]Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, Germany
{phan, hertel, maass, mazur, mertins}@isip.uni-luebeck.de

10.21437/Interspeech.2015-682

Abstract

The human auditory system is very well matched to both human speech and environmental sounds. Therefore, the question arises whether human speech material may provide useful information for training systems for analyzing nonspeech audio signals, such as in a recognition task. To find out how similar nonspeech signals are to speech, we measure the closeness between target nonspeech signals and different basis speech categories via a speech classification model. The speech similarities are finally employed as a descriptor to represent the target signal. We further show that a better descriptor can be obtained by learning to organize the speech categories hierarchically with a tree structure. We conduct experiments for the audio event analysis application by using speech words from the TIMIT dataset to learn the descriptors for the audio events of the Freiburg-106 dataset. Our results on the event recognition task outperform those achieved by the best system even though a simple *linear* classifier is used. Furthermore, integrating the learned descriptors as an additional source leads to improved performance.
Index Terms: feature learning, audio event, speech model

1. Introduction

Beside human speech, the most important audio signal, computational analysis of other nonspeech audio signals (e.g. music [1, 2], environmental sounds [3, 4]) is becoming more and more important [5]. In this domain, signal representation remains a fundamental problem for many other successive tasks such as recognition [1, 6] and detection [7, 2].

Many works have focused on the development of efficient signal representations. Various hand-crafted descriptors have been proposed. Most of them are borrowed from speech representations, such as mel-scale filter banks [8], log frequency filter banks [9], and time-frequency features [10, 11]. With the rapid advance of machine learning, automatic feature learning is becoming more and more common [12, 13, 14, 15]. Although considerable progress has been made in individual problems, more often than not, these representations are derived based on analysis of the target signals per se. We still lack a general way of representing audio signals and specifically lack a universal descriptor for them. Such a generic representation would be very helpful for solving various audio analysis tasks in a homogeneous way.

In this work, we propose such a generic descriptor for nonspeech audio signals by measuring the correlations between the target signal and different speech signals. The speech signals are obtained from an external source which is not related to the target audio signal of interest. To accomplish this, given a set of labeled speech signals of different categories (e.g. speech words), we are able to learn a multi-class speech classifier. Inputting the target signal into the speech classifier, we obtain the likelihoods that it is classified to different speech categories modeled by the classifier. These likelihoods can be interpreted as the acoustical closeness between the target signal and the basis speech signals. In intuition, they measure how the target signal sounds like the sounds of the speech signals. Eventually, we used the speech classifier as a feature extractor and the speech similarities are used to describe the target audio signal. The idea is illustrated in Figure 1. By collecting a sufficiently large set of basis speech categories, we are able to cover a wide range of acoustic concepts of the world. As a result, embedding the target audio signal into the space spanned by these bases is expected to produce a good representation. We will show that a better representation can be achieved by automatically constructing a label tree to organize the speech categories hierarchically and learn multiple speech classifiers for feature extraction along the tree accordingly. The proposed descriptors are generic in the sense that once the feature extractors are trained, they can be used to extract features for any input signals without re-training.

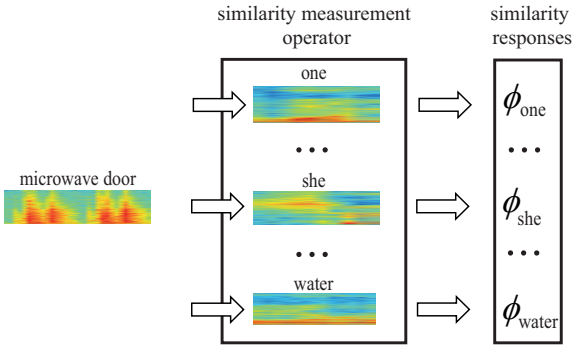

Figure 1: The "microwave door" audio event is represented by its similarities to different speech words such as "one", "she", and "water".

A few works have explored additional data sources (e.g. multiple channels [16], multiple modalities [17]) to augment the analysis. However, the main goal is to compensate for low signal-to-noise-ratio and overlapping signals. Therefore, not surprisingly, the additional data are of the same signal under analysis. Differently, our goal is to learn representations for a target audio signal via external speech signals which are totally unconnected to the target signal. In our experiments, we learn the descriptors for audio event signals of the Freiburg-106 dataset [18] through speech words of the TIMIT dataset [19]. We show that our event recognition systems outperform those achieved by the best system even though a simple linear classifier is used. Furthermore, fusing the learned descriptors as an additional source leads to improved performance of the system built on the audio signals themselves.

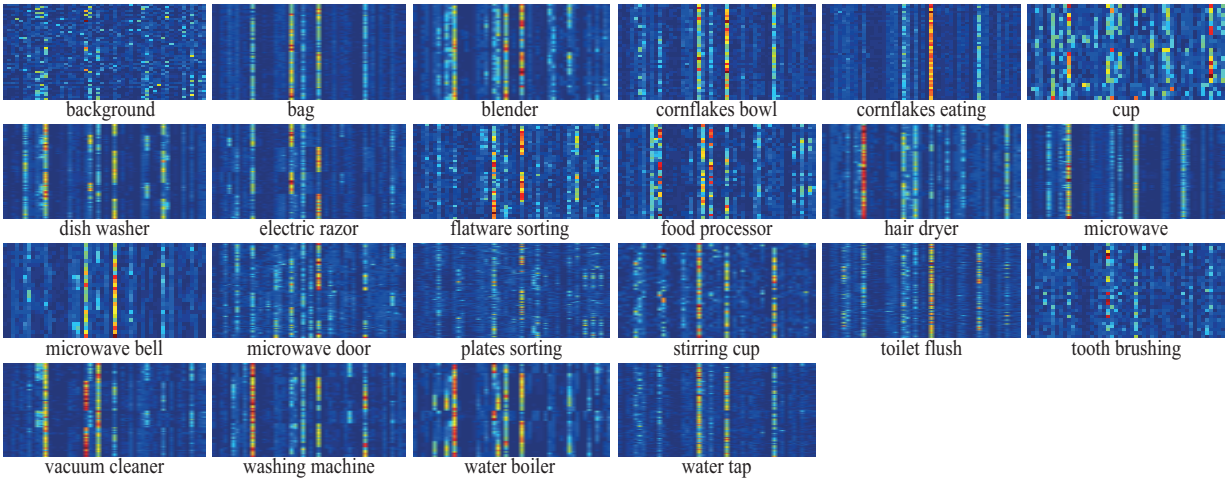

Figure 2: Similarities between audio events of the Freiburg-106 dataset and 50 speech word categories of the TIMIT dataset. Each row of the image represents one event of the corresponding class.

2. The approach

In the following, we propose two types of similarity descriptors. In Section 2.1 we look at descriptors that directly measure similarities between different categories. In Section 2.2 we then build tree-induced descriptors.

2.1. Nonspeech audio signal representations via speech similarities

Given a database of speech signals $\mathcal{S} = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$, where $\mathbf{x}_i$ denotes the low-level descriptor for the i-th signal (e.g. MFCCs [8] or log frequency filter bank parameters [9]) and $c_i \in \{1, \ldots, C\}$ indicates the class label. The C speech classes are used as our bases and they should ideally include all possible acoustic concepts.

Let us denote the target audio signal as $\mathbf{x}_e$. Our goal is to represent the target signal in terms of its acoustical closeness to the set of C basis speech classes. We accomplish this using some classification models. Intuitively, one can learn C 1-vs-the-rest binary classifiers each of which recognizes the c-th speech category. Such a classifier is trained using the c-th category as positive examples and the other $C-1$ classes as negative examples. Alternatively, for convenience, we jointly learn a multi-class speech classifier $\mathcal{M}_\mathcal{S}$ at once using random forest classification [20]. The target event $\mathbf{x}_e$ is then inputted into $\mathcal{M}_\mathcal{S}$ to obtain the classification posterior probabilities $\phi = [\phi_1, \ldots, \phi_C] \in \mathbb{R}_+^C$ where $\phi_c = P(c|\mathbf{x}_e)$ and $c \in \{1, \ldots, C\}$. Each entry ϕ_c quantifies how likely the target event belongs to the event category c of $\mathcal{S}$, i.e. it can be interpreted as a similarity measure.

Traditionally, the posterior probabilities produced by the classifier $\mathcal{M}_\mathcal{S}$ are used to make decisions, e.g. in a recognition task. In this work, we use the classifier $\mathcal{M}_\mathcal{S}$ as a feature extractor, and the vector ϕ is used as a descriptor for the event $\mathbf{x}_e$. As a result, the audio event is embedded in the space spanned by the speech similarities. In Figure 2, we illustrate the similarities of audio events in the Freiburg-106 dataset [18] to 50 speech word categories of TIMIT dataset [19]. The word categories were selected randomly and we trained the classifier $\mathcal{M}_\mathcal{S}$ with 200 trees. We can see distinguished patterns on different categories, for example "cornflakes eating". In particular, the "background" class shows random response since it contains different diverged sounds. Overall, the audio events are distinguishable by representations through the speech basis classes.

2.2. Learning a label tree of basis speech categories

We argue that in order to learn for good descriptors, we need to choose a set of varied speech categories. With expertise, one can carefully select such speech categories by hand. Here, we propose to discover them from a randomly pre-determined set $\mathcal{S}$. We collectively partition the speech categories into subsets such that they are easy to distinguish from one another. For this purpose, we learn a label tree for the speech categories similarly to [21]. This algorithm was originally proposed to learn a tree structure of classifiers (the label tree). Instead, we use it to form the sets of speech categories that can be easily distinguished.

Let $\ell_\mathcal{S} \equiv \{1, \ldots, C\}$ denote the label set of the speech database $\mathcal{S}$. The label tree is constructed recursively so that each node is associated with a set of class labels. Let us consider a node with a label set ℓ (and therefore, the root node is with the label set $\ell_\mathcal{S}$). We want to split the set ℓ into two subsets ℓ^L and ℓ^R where $\ell^L \neq \emptyset$, $\ell^R \neq \emptyset$, $\ell^L \cup \ell^R = \ell$, and $\ell^L \cap \ell^R = \emptyset$. There are totally $2^{|\ell|-1} - 1$ possible partitions $\{\ell^L, \ell^R\}$ where $|\cdot|$ denotes the cardinality. We want to select the partition such that a binary classifier to distinguish ℓ^L and ℓ^R makes as few errors as possible. The exhaustive search for such a partition would be prohibitively expensive especially when $|\ell_\mathcal{S}|$ is large. Instead, we rely on the confusion matrix of a multi-class classifier to determine a good partitioning. Our goal is to include classes that tend to be confused with each other in the same subset. Let $\mathcal{S}^\ell \subset \mathcal{S}$ denote the set of speech signals corresponding to the label set ℓ. Furthermore, suppose that we have changed and sorted the label set ℓ so that $\ell = \{1, \ldots, |\ell|\}$. To obtain the confusion matrix, we divide $\mathcal{S}^\ell$ into two halves: $\mathcal{S}^\ell_{train}$ to train the classifier and $\mathcal{S}^\ell_{val}$ for validation. Again, we train the multi-class classifier using random forest classification. Let $\mathbf{A} \in \mathbb{R}^{|\ell| \times |\ell|}$ denote the confusion matrix of the classification on the validation set $\mathcal{S}^\ell_{val}$. Each element $\mathbf{A}_{ij}$ is given by:

$$\mathbf{A}_{ij} = \frac{1}{|\mathcal{S}^\ell_{val,i}|} \sum_{\mathbf{x} \in \mathcal{S}^\ell_{val,i}} P(j|\mathbf{x}) \qquad (1)$$

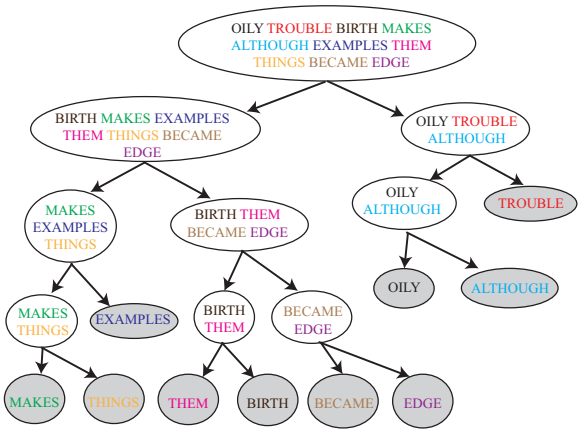

Figure 3: The learned label tree for 10 randomly selected TIMIT word categories. The white and shaded nodes represent the split and leaf nodes respectively.

where $\mathcal{S}^{\ell}_{val,i} \in \mathcal{S}^{\ell}_{val}$ are the speech signals with label i. $\mathbf{A}_{ij}$ expresses how likely a speech sample of class i is predicted to belong to class j by the classifier. Since $\mathbf{A}$ is not symmetric, we symmetrize it as

$$\bar{\mathbf{A}} = (\mathbf{A} + \mathbf{A}^T)/2. \tag{2}$$

Eventually, the optimal partitioning $\{\ell^L, \ell^R\}$ is selected to maximize:

$$E(l) = \sum_{i,j \in \ell^L} \bar{\mathbf{A}}_{ij} + \sum_{m,n \in \ell^R} \bar{\mathbf{A}}_{mn}. \tag{3}$$

By this, we tend to group the ambiguous speech categories into the same subset, as a result, produce two meta-classes $\{\ell^L, \ell^R\}$ that are easy to separate from each other. We apply spectral clustering [22] on the matrix $\bar{\mathbf{A}}$ to solve (3).

Once the optimal partition $\{\ell^L, \ell^R\}$ is determined, we learn another binary classifier $\mathcal{M}^{\ell}_{\mathcal{S}}$. We use the set S^{ℓ} as training data. The samples with their labels in ℓ^L are considered as negative examples and others with their labels in ℓ^R are considered as positive examples. The classifier $\mathcal{M}^{\ell}_{\mathcal{S}}$ is then associated with the node and used as a basis classifier for feature extraction. We recursively repeat the process until a single class label remains at a node.

This procedure produces totally $|\ell_{\mathcal{S}}| - 1$ basis classifiers associated with the split nodes of the tree. Evaluating them on the target audio signal $\mathbf{x}_e$ will produce a feature vector of size $2(|\ell_{\mathcal{S}}| - 1)$ to describe it. It is noticed that the tree construction and evaluation can be done in parallel, therefore, it is computationally efficient. In Figure 3, we show a label tree constructed for ten randomly selected speech word categories of the TIMIT dataset using the algorithm. Note that, unlike WordNet [23], this tree does not need to capture any semantic of the words.

3. Experiments

The descriptors derived in Section 2 are generic rather than specific for a certain application. That is, once the feature extractors are learned, we can use them to extract representations for any inputted audio signal such as music, audio events, etc. They are different from other features learned by a conventional way, such as bag-of-words representations [24, 13, 14], which are task-specific and data-specific.

3.1. Experimental setup

Test datasets. We used the Freiburg-106 dataset [18] and TIMIT dataset [19] to test our approach. The Freiburg-106 audio events are considered as nonspeech target signals, and the basis speech categories were extracted from the TIMIT dataset.

The Freiburg-106 dataset was collected using a consumer-level dynamic cardioid microphone. It contains 1,479 audio-based human activities of 22 categories. As in [18], we divided the dataset so that the test set contains every second recording of a category and the training set contains all the remaining recordings[1].

Using the TIMIT speech database, different representation levels (e.g. phonemes, words, and sentences) may be considered. To demonstrate the proposed concept, we use word categories here. We randomly selected C word categories with $C = \{50, 100, \ldots, 500\}$ for the experiments. Only speech words that occur more than ten times in the dataset were used, and we only kept at most 50 samples per class.

Low-level features to represent a signal. The signals (i.e. audio events and speech words) were firstly downsampled to 16 kHz. Each audio event was decomposed into 50 ms segments with a step size of 10 ms. Whereas, those used for a speech signal were 25 ms and 10 ms respectively as usual use for speech. A longer segment size was used for audio event to better capture their nonstationary effects [7].

Although any arbitrary low-level features are feasible to describe a segment, we extracted a set of very basic acoustic features for every audio segment: 16 log-frequency filter bank coefficients [9], their first and second derivatives, zero-crossing rate, short-time energy, four sub-band energies, spectral centroid, and spectral bandwidth. Totally, there were 53 features for each segment. In turn, a whole signal is represented by the 106-dimension feature vector computed the mean and standard deviation over its segments.

Other parameters. For the random forest classifiers used in Sections 2.1 and 2.2, we trained them with the algorithm in [20] with 200 trees each.

Audio event classification models. We trained our event classification systems using one-vs-one SVMs with different kernels, including linear, RBF, χ^2, and histogram intersection (hist. for short). Except for the RBF kernel, the hyperparameters C of the SVMs were tuned via leave-one-out cross-validation. For the one with the RBF kernel which is usually computationally expensive , we conducted 10-fold cross-validation to search for the hyperparameters and the kernel parameters.

Evaluation metrics. For evaluation of classification performance, we make use of the *f-score* metric, which considers both precision and recall values:

$$\text{f-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{4}$$

3.2. Experimental results

Flat descriptors vs. tree-induced descriptors. Let us denote the descriptors described in Section 2.1 as flat descriptors opposing to the tree-induced descriptors in Section 2.2.

The performance of these two descriptors for the audio event recognition task is shown in Figure 4. Obviously, with the same speech bases, the tree-induced descriptors perform much better than the flat counterparts. Specifically, the average improvements are 5.87%, 5.69%, 3.45%, and 4.81% with respec-

[1] This is based on unofficial communication with the authors of [18].

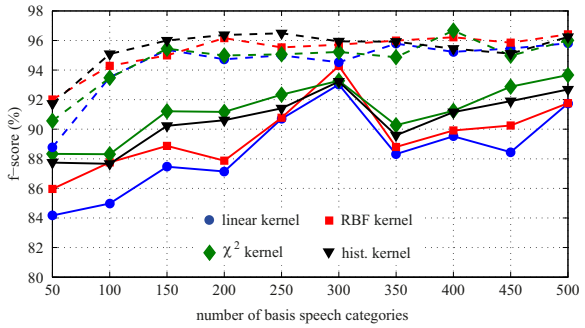

Figure 4: Performance of the flat descriptors (solid lines) and the tree-induced descriptors (dash lines) on audio event recognition with different kernels.

Table 1: Average performances on f-score (%) of the proposed descriptors compared to the state-of-the-art on the Freiburg-106 dataset (92.4% [18]).

Descriptors	Linear	RBF	χ^2	Hist.
Flat	88.55	89.63	91.27	90.62
Tree-induced	94.42	95.31	94.73	95.43

tive to linear, RBF, χ^2, and hist. kernels. It is also worth noticing that the performance of the linear classifiers are comparable with the other nonlinear classifiers while they are computationally much cheaper to train and evaluate.

Compared to the state-of-the-art performance on the Freiburg-106 dataset (92.4% on f-score [18]), the average performances of our systems are shown in Table 1. While the flat descriptors underperform, the tree-induced descriptors outrun the state-of-the-art even with a simple linear classifier. These results are impressive given the fact that we have not used the low-level features of the audio events in the models.

Using the descriptors as additional features. In this experiment, we studied how the proposed descriptors improve the recognition with some fusion schemes when we considered them as additional features. We implemented a bag-of-words (BoW) model, which has been widely used for the audio event recognition task [24, 13, 25, 14], using low-level frame-based features of the audio events.

We used k-means for codebook learning. The entries were obtained as the cluster centroids, and codebook matching was based on Euclidean distance. After obtaining BoW representations, the classifier was learned using SVM with a χ^2 kernel. Again, the hyperparameters were tuned via leave-one-out cross validation. Since the performance of such BoW models heavily depends on the codebook size, we conduct the analysis with different codebook sizes $\{50, 75, \ldots, 250\}$.

Different descriptors (i.e. the BoW descriptors and the proposed descriptors) are then combined in a simple multi-channel approach [26]:

$$K(e_i, e_j) = \exp\big(-\sum_k \frac{1}{M^k} D(e_i^k, e_j^k)\big) \qquad (5)$$

where $D(e_i^k, e_j^k)$ is the χ^2 distance between the audio events e_i and e_j with respect to the k-th channel. M^k is the mean χ^2 distance of the training samples for the k-th channel. For classification, we used a nonlinear SVM with an RBF-χ^2 kernel [27].

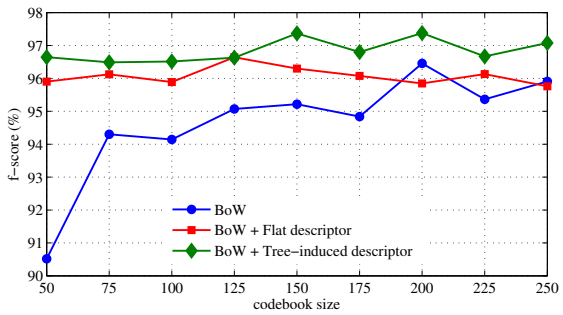

Figure 5: Recognition performance by fusing the proposed descriptors with BoW descriptors.

The fusion results are shown in Figure 5. The fusion systems lead to 1.43% and 2.19% average improvement with the flat and tree-induced descriptors, respectively, compared to the BoW descriptors.

4. Discussion

The fact is that more than 6900 languages in the world [28] and many annotated corpuses are available such as TIMIT [19] and GlobalPhone [29]. It opens enormous opportunities to explore for learning representations from speech. Using different levels (e.g. phonemes, words) and different languages would result in different representations. Their combinations would offer even more opportunities.

It can be seen from Figure 4 that the number of basis speech categories needs to be sufficiently large to guarantee a good performance. This is understandable since with more basis speech, we are likely to cover more acoustic concepts. However, just increasing the number of bases does not guarantee a better performance. The reason is quite obvious. For example, when the bases are randomly selected, many similar categories (e.g. "become", and "becomes") are likely to exist. This results in correlation in some dimensions of the induced feature space which worsen the model. As shown, organizing the bases in a tree structure is efficient to alleviate this problem. However, it is worth further studying how to deal with it.

5. Conclusions

We present in this paper the idea to represent a target nonspeech audio signal by its similarities to different basis speech signals. We further proposed to learn to organize the basis speech categories within a tree structure to achieve a better representation. Our experiments on the audio event recognition task show that the proposed descriptors are efficient even with a simple linear classification model. They can also act as additional features to augment an existing system to obtain a better performance. The use of the word level was quite arbitrary in our study. Further work will be directed toward defining optimally suited categories, for example, in form of triphones and other speech segments.

6. Acknowledgements

This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. We would also like to thank Johannes A. Stork for providing the Freiburg-106 dataset.

7. References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 291–301, 2008.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1994.

[4] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York: IEEE Press, 2006.

[5] R. F. Lyon, "Machine hearing: An emerging field," *Signal Processing Magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[6] H. Phan and A. Mertins, "Exploring superframe co-occurrence for acoustic event recognition," in *Proc. EUSIPCO*, 2014, pp. 631–635.

[7] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[9] C. Nadeu, D. Macho, and J. Hernando, "Frequency and time filtering of filter-bank energies for robust hmm speech recognition," *Speech Communication*, vol. 34, pp. 93–114, 2001.

[10] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 367–377, 2013.

[11] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[12] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, 2009, pp. 1096–1104.

[13] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *Proc. ICASSP*, 2014, pp. 1370–1374.

[14] A. Plinge, R. Grzeszick, and G. Fink, "A bag-of-features approach to acoustic event detection," in *Proc. ICASSP*, 2014, pp. 3704–3708.

[15] E. Humphrey, J. Bello, and Y. Lecun, "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics," in *Proc. ISMIR*, 2012, pp. 403–408.

[16] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *EUSIPCO 2014*, 2014, pp. 2375–2379.

[17] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, and J. R. Casas, "Acoustic event detection based on feature-level fusion of audio and video modalities," *EURASIP Journal on Advances in Signal Processing*, 2011.

[18] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-Markovian ensemble voting," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'12)*, 2012, pp. 509–514.

[19] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[20] L. Breiman, "Random forest," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] S. Bengio, J. Weston, and D. Grangier, "Label embedding trees for large multi-class tasks," in *Proc. NIPS*, 2010, pp. 163–171.

[22] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS*, 2001, pp. 849–856.

[23] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller, "Introduction to wordnet: An online lexical database," *International Journal of Lexicograph*, vol. 3, no. 4, pp. 235–244, 1990.

[24] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. Interspeech*, 2013.

[25] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2013, pp. 81–86.

[26] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[27] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc CVPR*, 2008, pp. 1–8.

[28] R. Gordon, Ed., *Ethnologue: Languages of the World*. Dallas: SIL International, 2005.

[29] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *Proc. ICASSP*, 2013, pp. 8126–8130.