International Journal on Artificial Intelligence Tools
Vol. 26, No. 2 (2017) 1750005 (13 pages)
© The Author(s)
DOI: 10.1142/S0218213017500051

Data-Driven Audio Feature Space Clustering for Automatic Sound Recognition in Radio Broadcast News

Theodoros Theodorou
Artificial Intelligence Group, Wire Communications Laboratory
Department of Electrical and Computer Engineering, University of Patras
26500 Rion-Patras, Greece
theodorou@upatras.gr

Iosif Mporas
School of Engineering and Technology, University of Hertfordshire, College Lane Campus
Hatfield AL10 9AB, Hertfordshire, United Kingdom
i.mporas@herts.ac.uk

Alexandros Lazaridis
Idiap Research Institute, Martigny, Switzerland
alaza@idiap.ch

Nikos Fakotakis
Artificial Intelligence Group, Wire Communications Laboratory
Department of Electrical and Computer Engineering, University of Patras
26500 Rion-Patras, Greece
fakotaki@upatras.gr

Received 3 February 2015
Accepted 17 November 2016
Published 7 April 2017

Aiming to an automatic sound recognizer for radio broadcasting events, a methodology of clustering the audio feature space using the discrimination ability of the audio descriptors as a criterion, is investigated in this work. From a given and close set of audio events, commonly found in broadcast news transmissions, a large set of audio descriptors is extracted and their data-driven ranking of relevance is clustered, providing a more robust feature selection. The clusters of the feature space are feeding machine learning algorithms implemented as classification models during the experimental evaluation. This methodology showed that support vector machines provide significantly good results, considering the achieved accuracy due to their ability of coping well in high dimensionality experimental conditions.

Keywords: Sound recognition; audio features; feature subspace selection.

1. Introduction

Over the last decade, there is consecutive increase of the available data accessible by an increasing number of people. Radio and TV broadcast transmissions and the web-based multimedia data offer an enormous amount of audiovisual data. The availability of these resources has led research to focus on a vast number of applications related to automatic processing of such multimedia data including TV program automatic handling, story classification, automatic highlighting of events, sports news handling, automatic transcription extraction, automatic commercial detection, summarization etc.[1–9]

Concerning the audio data, the automatic analysis of the audio signals can offer the users useful information. In the case of broadcast news, automatic processing is related to tasks such as sound recognition,[10,11] speaker recognition,[12] anchor detection,[13] role detection,[14–16] story boundary detection,[2,17,18] summary construction from anchor talking,[9,19] channel's quality detection,[20] sound event detection,[21,22] non-linguistic human-produced sounds detection,[5,6,23–25] audio type segmentation in sport games,[4,26,27] highlight scene extraction from sports games,[3] violence scene detection,[28] music characteristics classification,[29,30] jingle detection,[1] commercial block detection,[8] voice activity detection,[31] language recognition,[32] emotion recognition[33] and speech recognition.[34] Sound recognition is the cornerstone of analysis as typically precedes the other stages.

During sound recognition the audio signal is decomposed to discrete intervals corresponding to sound events of interest. In broadcast news signals additionally to the major sound categories of the speech and music, common sounds are the non-linguistic sounds, noise from the recording/transmission conditions, bubble noise, background/ environmental sounds and superposition of sounds. For the decomposition of the broadcast signal, the signal is initially preprocessed and parameterized. Consequently, the parameterized signal is processed by a pattern recognition algorithm.

Over the years, several time-domain and frequency-domain features have been used for parameterizing the broadcast audio signals.[10,35,36] Zero crossing rate and the Mel frequency cepstral coefficients are the most commonly used in time-domain and in the frequency-domain correspondingly. Other commonly used features are the pitch, perceptual linear predictive coefficients, harmonics-to-noise ratio, linear predictive coding coefficients, chroma, autocorrelation etc.[10,26,35,36,38,39]

In the pattern recognition stage, a big variety of probabilistic and discriminative machine learning algorithms have been proposed. The most commonly used are the Gaussian mixture models and the hidden Markov models.[10,11,14,26,37,40] Also widely used are the support vector machines,[11,14,38,39,41] the artificial neural networks,[10] the k-nearest neighbor algorithm,[14,38] the decision trees,[10,38] the genetic algorithms,[2] the fuzzy logic[42] and boosting techniques.[41,43]

Related architectures incorporate fusion frameworks among recognition models[28, 44] and combination of model-based and distance based algorithms.[13,26,27,39,40] Post-processing schemes can improve the overall recognition accuracy. Among the post-processing schemes are (i) transformation of the feature matrix,[23,44–46] (ii) correction of logical errors based on empirical rules,[11] (iii) isolation of the segments of interest in cases where the post-processing is focused on specific classes[10,11,13,38,40,47] and (iv) merging of sound events and separation of them in a post-processing stage.[28]

The structure of the analysis of sounds categorizes the task to different classes. Some of the widely used are: (i) multi-class problem,[10] (ii) binary-classes problem,[37] (iii) hierarchical structure of the classes problem,[11] (iv) two-groups or multi-group of classes problem[28] and (v) detection of a class over the other classes problem.[19,48]

In this work, we present a broadcast news sound recognition methodology based on widely known and used audio features. The implemented framework clusters the audio feature space to subspaces, based on data-driven criteria. Consequently, the subspaces that are found useful in terms of their sound discrimination ability are utilized in the sound recognition task. We concentrate our interest in investigating our methodology based on main hypotheses that are expected to be verified. The first hypothesis is that clustering the audio feature space using the discrimination ability of the audio descriptors as a criterion will be beneficial to the task. Secondly, even though most of the machine learning algorithms incorporate the ability of identifying the most appropriate features and discard the rest, the use of irrelevant features often deteriorates the effectiveness of the algorithms. Also we hypothesize that clustering the features using the discrimination ability not only avoids deterioration but also incorporates a more robust stage between feature extraction and recognition that assists the methodology. Finally, algorithms able to cope well with high dimensional feature spaces, like the SVMs, will manage to perform very well in all models. In this way, it is expected to achieve the optimization of the classification accuracy, avoiding a time consuming greedy feature selection approach.

The rest of the article is organized as follows: In Section 2, the proposed methodology for recognition of sounds using clustering of the audio feature space is described. In Section 3, the experimental setup is given and in Section 4, the experimental results are presented. Finally, in Section 5 we conclude this work.

2. Sound Recognition with Unsupervised Audio Feature Space Clustering

In the proposed scheme the recognition of the sounds of interest is based on short-time analysis in the time and frequency domain. A selection of clusters of the feature space, that are expected to be more discriminative with respect to the sounds of interest, is implemented. Figure 1 illustrates this framework. As shown, the proposed scheme is divided into two phases, the training and test phase.

During the training phase a set of R audio recordings $X = \{X^r\}$, $1 \le r \le R$, with known sound labels, is used to train models for each of the sound types of interest. The training phase consists of the pre-processing, feature extraction, evaluation of features for clustering and classification model construction steps. During pre-processing the training audio files, $X = \{X^r\}$, are frame blocked with overlapping frames, $O = \{O^r\}$ of constant length with constant time-shift step. The sequences of audio frames are decomposed to sequences of feature vectors, $V = \{V^r\}$, in the feature extraction block. The feature extraction block applies a number of feature extraction algorithms to each audio frame

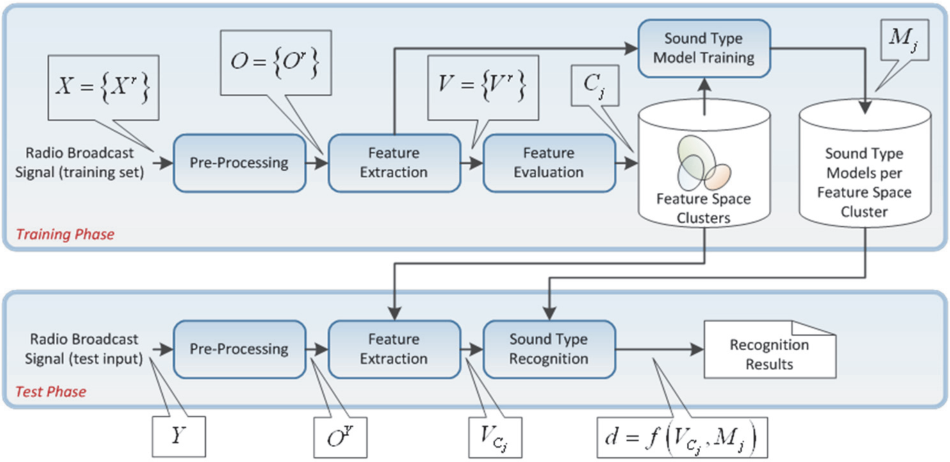

Fig. 1. General architecture of the proposed scheme.

and the computed feature values are concatenated in one feature vector, $V_i^r \in R^K$, per audio frame *i*. After decomposing the audio signals to sequences of feature vectors, a ranking algorithm is applied by the feature evaluation block. The output of this block is a number of feature clusters, C_j, with $1 \le j \le J$, which divide the feature space to *J* clusters with respect to the estimated ranking score, i.e. the discriminative ability of the features. Consequently cluster C_1 will include the most discriminative features, while cluster C_J will include the less discriminative ones. The number of clusters *J* is either manually defined or determined by a threshold criterion with respect to the sparseness of each cluster. The clustering of the feature space allows the training of sound type models with subsets of the features instead of using the entire feature space. During the training step of the classification model, the *j* most valuable clusters are used to train model M_j, i.e. model M_1 is trained with the features of cluster C_1 and so on until the last model M_J will be trained with all clusters. The training phase results in *J* models for radio broadcast sound type classification.

During the test phase, an unknown test audio file, *Y*, is pre-processed similarly to the training phase, i.e. with the same frame length and time-shift step, resulting in a sequence of frames, O^Y. The pre-processed audio signal, O^Y, is then processed by the feature extraction block estimating those features, V_{C_i}, that belong to a number of selected clusters C_j. The selection of the clusters is performed manually depending on the experimental setup. The estimated sub-feature sequence, V_{C_i}, is then forwarded to the classifier *f*, where a decision *d* is taken with respect to the corresponding model M_j of the selected features, i.e. $d = f(V_{C_i}, M_j)$. The recognition is based on frame-level classification among a closed set of sound types. Further post-processing of the results can be performed for fine-tuning of the estimated sound type intervals. The described architecture allows the exploitation of the feature subspaces, which contribute to the robust discrimination, excluding the feature subspaces that do not contribute.

3. Experimental Setup

The experimental setup for the evaluation of the architecture described in Section 2, is presented here. In this framework we are interested in examining our methodology validating the main hypotheses mentioned in the Introduction. As SVMs are able to cope well in high dimensional feature space, it is expected that will manage to perform excellent in comparison with all other models and probably outperform them. Clustering the audio feature space using the discrimination ability of the audio descriptors as a criterion, will be beneficial in the task. The audio data used for the evaluation, the feature extraction algorithms and the classification methods are also described in this section.

3.1. *Audio data description*

For our task, due to the lack of one database appropriate for sound type recognition from broadcast recordings, we relied on a number of existing audio data collections. The data collections used are (i) the Voice of America VOA radio broadcast news[49] for the Greek language, which is part of the NIST 2009 Language Recognition Evaluation,[50] (ii) the BBC FX Library,[51] (iii) the BBC broadcast news database,[52] (iv) the Partners In Rhyme database[53] and (v) the SoundBible database.[54] Sound instances acquired from non-broadcast collections were convolved with randomly selected silence intervals from broadcast audio signals. All audio data were stored in single-channel audio files with sampling frequency 8 kHz and resolution analysis 8 bits per sample. The selected audio data collection consists of recordings with total duration of approximately 8 min. The duration distribution per sound type is illustrated in Table 1.

The collected data include the most common sounds found in radio broadcasts. The entire evaluation audio dataset was manually annotated by an expert audio engineer.

Table 1. Duration distribution of the sound types in the collected audio data.

Sound	Duration (sec.)	Sound	Duration (sec.)	Sound	Duration (sec.)
Applause	58.86	Laugh	65.24	Silence	34.73
Bubble Noise	68.18	Music	94.73	Speech	98.32
Cough	43.11				

3.2. *Feature extraction*

The sound types appearing in radio broadcast signals differ in kind (speech, music, etc.). In the literature most of the feature extraction algorithms are dedicated to specific audio signals, mainly speech and music. In this study, we rely on the OpenSmile[35] framework for extracting a number of features that have been widely used in applications related to speech, music and sound recognition. The audio signal is initially frame blocked to overlapping frames of constant length of 25 msec. A 1st order FIR pre-emphasis filter followed by Hamming windowing is applied to each frame. From each frame we compute (i) the zero-crossing rate, (ii) the frame energy, and after computing the spectral magnitude we compute (iii) the Mel frequency cepstral coefficients,[55] (iv) the pitch envelope, (v) the voice probability, (vi) the chroma coefficients,[56,57] and the spectral magnitude statistics (vii) energy per 4 equally distributed at 0-$F_S/2$ bands, (viii) roll off, (ix) flux, (x) centroid, frequency with (xi) maximum and (xii) minimum magnitude. All audio features are concatenated to a common feature vector, which is further expanded with first and second derivatives (delta and delta-delta coefficients).

3.3. *Feature evaluation and clustering*

After computing the audio features described in subsection 3.2, the feature evaluation block estimates the importance of each feature, with respect to their discriminative ability on the task. For the evaluation we relied on the ReliefF algorithm.[58] The ReliefF algorithm computes a vector W of the estimations of the qualities of all the audio features. The ranking position of each feature is defined by its ranking score, i.e. the corresponding estimation of quality, $w \in R$, which indicates the degree of importance of that feature. These ranking scores are used to cluster the feature space into five clusters using the EM algorithm.[59] The number of the clusters was chosen based on empirical knowledge and preliminary experiments. In detail, the ranking scores, $w \in R$, are used to iteratively train five one-dimensional Gaussian distributions, each for one cluster. After the completion of the EM training each feature is assigned to the cluster where the ranking score has the maximum likelihood. The usage of EM ensures the maximum likelihood in the distribution models. This clustering procedure ensures that attributes with close ranking scores will be grouped together in the same cluster, since their importance is alike, resulting clusters corresponding to meaningful subsets of features. The clusters are used for estimating classification models with different sets of clusters during the training phase and for cluster-specific feature extraction during the test phase.

3.4. *Sound type classification*

For the construction of the classification models we used the implementations of machine learning algorithms of the WEKA software toolkit.[60] Well-known and widely used, in the areas of audio, speech and music processing, algorithms were selected.[10,11,38] The evaluated algorithms are: (i) a two-layered back-propagation multilayer perceptron (MLP) neural network,[61] (ii) a support vector classifier (SVM) with radial basis function kernel utilizing the sequential minimal optimization algorithm,[62] (iii) a k-nearest neighbor classifier (IBk)[63] and (iv) a C4.5 decision tree learner (J48).[64] The hyper-parameters of all algorithms were selected using grid search. For the purpose of direct comparison, all algorithms were trained with the same training data and evaluated on the same test data. For each cluster combination, one model was trained.

4. Experimental Results

The architecture presented in Section 2 was evaluated using the experimental protocol described in Section 3. The performance of the four algorithms was evaluated on frame level. In order to avoid overlap between the training and test subsets a ten-fold cross validation experimental setup was followed. The achieved results for the full audio feature vector are shown in Table 2.

Table 2. Broadcast news sound recognition accuracy for different algorithms.

Algorithm	MLP	SVM ($C = 5$, $g = 0.01$)	IBk ($k = 2$)	J48
Accuracy (%)	91.27	96.02	95.17	94.03

As can be seen in Table 2, the SVM classification algorithm outperformed all the other algorithms achieving accuracy of 96.02%. The second-best performing algorithm was the IBk, which achieved approximately 1% lower performance. Both the decision tree and the neural network achieved significantly lower performance. The advantage of the SVM algorithm can be explained by the ability of SVMs to cope well with the high dimensionality of the feature space in respect to the amount of data, since they do not suffer from the curse of dimensionality[62] and, in contrast to the rest algorithms, will converge to the global optimal parameter values, and thus will not provide suboptimal performance. In a further step we evaluated the discriminative ability of each feature in order to investigate the effect of dimensionality reduction. The choice of five clusters was empirically, without this decision undermining criteria-based decisions that respects the likelihood of the data. The resulting feature subset clusters consisted of 16, 18, 12, 71 and 90 features respectively. In Table 3, we present the audio features of the 1st cluster. The selection of clusters was defined during the training phase.

Table 3. Top-16 audio features (assigned to cluster 1) according to the ReliefF criterion.

Ranking	Feature	Ranking	Feature	Ranking	Feature
1	Pitch envelope	7	minimum value	12	Centroid
2	12th MFCC	8	Zero Crossing Rate	13	Pitch
3	Energy	9	3rd MFCC	14	5th MFCC
4	0th MFCC	10	2nd MFCC	15	4th MFCC
5	1st MFCC	11	maximum value	16	RollOff-90%
6	Voicing Prob.				

As can be seen in Table 3, within the most discriminative features are the pitch (absolute value and envelope), several MFCCs, the energy, the voicing probability, some spectral magnitude statistics and the zero-crossing rate. These results are in agreement with Refs. 23 and 65, where the MFCCs, the zero crossing rate, the voicing probability and the pitch were found as discriminative features. In Table 4, we present the accuracy for all clusters and for each algorithm. The performance of each method for the full

Table 4. Accuracy for different audio feature subsets and classification algorithms.

Dimensionality Classifiers	$16/C_1$	$34/C_{1,2}$	$46/C_{1,2,3}$	$117/\mathrm{C}_{1,2,3,4}$	$207/\mathrm{C}_{1,2,3,4,5}$
MLP	86.39	87.96	90.04	92.45	91.27
SVM	89.44	90.68	94.29	96.37	96.02
IBk	89.05	90.42	94.95	95.33	95.17
J48	87.24	88.95	92.73	94.73	94.03

feature vector (i.e. $C_{1,2,3,4,5}$), showed in Table 2, is repeated here for the convenience of comparison.

As can be seen in Table 4, the use of subsets of features improves the overall performance. Specifically, the 117 best features in terms of discriminative ability ranking, which correspond to the 4-best clusters i.e. 43.5% reduction of the number of features used, compared to the full feature set, achieved the highest performance for all the evaluated algorithms. These results show that clustering the features by using the ranking is an effective method to discard irrelevant features and works in favor of the outcome even though most of the machine learning algorithms have the ability to learn which are the most appropriate features. This is owed to the significant reduction of the feature space dimension, which reduces the effect of the curse of dimensionality phenomenon as well as to the fact that the use of fewer features prevents from overfitting. For all sets of clusters the accuracy of the four classifiers is similar to the full feature set, i.e. the SVM algorithm outperforms all the other algorithms and is followed by the IBk algorithm. The main advantage of SVM that leads to outperforming all the other models is its fundamental property of coping well with high dimensional feature space[66] along with their ability to learn complex relationship between the input and output of the data.[67] Moreover it can be pointed out that in all cases, IBk managed to achieve performance close to SVM. In one case, i.e. $C_{1,2,3}$, IBk even managed to slightly (approximately 0.5%) outperform the SVM and give the highest performance in this set of clusters. On the contrary, when the dimensionality of the feature space increases a lot, i.e. the number of the features of $C_{1,2,3,4}$ and $C_{1,2,3,4,5}$ become 2.5 and 4.5 times the dimension of $C_{1,2,3}$ respectively, the performance of IBk deteriorates, since due to the large number of features, all data vectors are almost equidistant to the search query vector based on the Euclidean distance.[68] Finally, in no case, neither MLP nor J48 managed to achieve a performance close to SVM or IBk models. In the case of MLP, this could be attributed to the amount of available training data in respect to the feature space[69,70] and in the case of J48, to the over-fitting of the model to the training data.[64] The accession of the 5th cluster, in all algorithms, showed that the growth of the parameterization reserved to decrease the achieved accuracy. The small changes in accuracy after the addition of the last clusters shows the efficiency of choosing feature by clustering. In a further step, the confusion matrix of the SVM model for the case of $C_{1,2,3,4}$

Table 5. Accuracy confusion matrix (in percentages) for $C_{1,2,3,4}$ feature set.

Recogn. As →	Applause	Bubble N.	Cough	Laugh	Music	Silence	Speech
applause	99.77	0.00	0.20	0.03	0.00	0.00	0.00
bubble noise	0.01	93.11	0.09	0.09	1.30	0.00	5.40
cough	1.55	2.11	85.59	7.05	2.25	0.19	1.26
laugh	0.23	1.17	2.28	91.37	0.97	1.16	2.82
music	0.00	0.70	0.52	0.32	96.19	0.00	2.27
silence	0.00	0.07	0.07	1.32	0.00	97.56	0.35
speech	0.00	0.34	2.83	0.37	2.20	0.21	94.05

feature set case was calculated and is shown in Table 5. As can be seen in the table, applause, music and silence are the types that achieved the highest recognition accuracies, showing rates above 96%. Speech, bubble-noise, laugh and especially cough presented deterioration in their accuracy rates achieving recognition rates between 85.59% and 94.05%.

5. Conclusions

The development of automatic event processing is driven by the availability of events and the quantity of applications. Since automatic audio recognizers have been cornerstones in audio event procedures, several methodologies have been investigated. In the present work, we studied an automatic sound recognition framework based on short time analysis of audio events commonly found in radio broadcast transmissions. The set of audio descriptors were organized into clusters based on their discrimination ability, incorporating a more robust method of selection. Several well-known and widely used machine learning algorithms were used. SVM managed to outperform due to its ability to cope well in high dimensionality problems. A t-test showed that SVM offered statistically significant better results than the rest. The IBk gave high accuracies, due to the nature of the examining audio data set. The addition of clusters with less significant features showed that it does not reserve the maximum accuracy, while it can reverse the opposite.

References

1. J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins and D. Caseiro, Broadcast news subtitling system in Portuguese, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP* 2008) (March 31–April 4, 2008), pp. 1561–1564.
2. C.-H. Wu and C.-H. Hsieh, Story segmentation and topic classification of broadcast news via a topic-based segmental model and genetic algorithm, *IEEE Transactions on Audio, Speech and Language Processing* **17**(8) (November 2009) 1612–1623.
3. Y. Itoh, S. Sakaki, K. Kojima and M. Ishigame, Highlight scene extraction of sports broadcasts using sports news programs, *IEEE 10th Workshop on Multimedia Signal Processing* (*MMSP 2008*) (October 8–10, 2008), pp. 646–649.

4. M. Baillie and J.M. Jose, An Audio-based sports video segmentation and event detection algorithm, *Conference on Computer Vision and Pattern Recognition Workshop* (*CVPRW '04*) (June 27–July 2, 2004), p. 110.
5. Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, A. Smailagic, I. Lane and P. Zhang, CoughLoc: Location-aware indoor acoustic sensing for non-intrusive cough detection, *International Workshop on Emerging Mobile Sensing Technologies, Systems, and Applications* (Mobisense, San Francisco, CA, June 2011).
6. S. Matos, S. S. Birring, I. D. Pavord and D. H. Evans, Detection of cough signals in continuous audio recordings using hidden Markov models, *IEEE Transactions on Biomedical Engineering* **53**(6) (June 2006) 1078–1083.
7. J. Chaloupka, Design of audio-visual TV broadcast news transcription system prototype, *53rd International Symposium ELMAR-2011* (September 14–16, 2011), pp. 209–212.
8. N. Liu, Y. Zhao, Z. Zhu and H. Lu, Exploiting visual-audio-textual characteristics for automatic TV commercial block detection and segmentation, *IEEE Transactions on Multimedia* **13**(5) (October 2011) 961–973.
9. Y.-T. Chen, B. Chen and H.-M. Wang, A probabilistic generative framework for extractive broadcast news speech summarization, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(1) (January 2009) 95–106.
10. T. Butko and C. Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: Overview, results and discussion, *EURASIP Journal on Audio, Speech and Music Processing* **2011**(1) (2011).
11. E. Dogan, M. Sert and A. Yazici, Content-based classification and segmentation of mixed-type audio by using MPEG-7 features, *First Int. Conf. on Advances in Multimedia* (*MMEDIA '09*) (July 20–25, 2009), pp. 152–157.
12. M. Kotti, V. Moschou and C. Kotropoulos, Speaker segmentation and clustering, *Signal Processing* **88**(5) (May 2008) 1091–1124.
13. C. Delphine, Model-free anchor speaker turn detection for automatic chapter generation in broadcast news, *IEEE Int. Conf. on Acoustics Speech and Signal Processing* (*ICASSP 2010*) (March 14–19, 2010), pp. 4966–4969.
14. B. Bigot, I. Ferrane and J. Pinquier, Exploiting speaker segmentations for automatic role detection. An application to broadcast news documents, *Int. Workshop on Content-Based Multimedia Indexing* (*CBMI 2010*) (June 23–25, 2010), pp. 1–6.
15. G. Damnati and D. Charlet, Robust speaker turn role labeling of TV broadcast news shows, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP 2011*) (May 22–27, 2011), pp. 5684–5687.
16. W. Wang, S. Yaman, K. Precoda and C. Richey, Automatic identification of speaker role and agreement/disagreement in broadcast conversation, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP 2011*) (2011), pp. 5556–5559.
17. L. Xie, Y. Yang, Z.-Q. Liu, W. Feng and Z. Liu, Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using maximum entropy approach, *Int. Conf. on Audio Language and Image Processing* (*ICALIP 2010*) (November 23–25, 2010), pp. 407– 413.
18. M.-M. Lu, L. Xie, Z.-H. Fu, D.-M. Jiamg and Y.-N. Zhang, Multi-modal feature integration for story boundary detection in broadcast news, *7th Int. Symp. on Chinese Spoken Language Processing* (*ISCSLP 2010*) (November 29–December 3, 2010), pp. 420–425.
19. S. H. Yella, V. Varma and K. Prahallad, Significance of anchor speaker segments for constructing extractive audio summaries of broadcast news, *IEEE Spoken Language Technology Workshop* (*SLT 2010*) (December 12–15, 2010), pp. 13–18.
20. M. Cettolo, Segmentation, classification and clustering of an Italian broadcast news corpus (*RIAO 2000*) (Paris, France, April 12–14, 2000), pp. 372–381.

21. H. D. Tran and H. Li, Sound event recognition with probabilistic distance SVMs, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(6) (August 2011) 1556–1568.
22. M. Wollmer, E. Marchi, S. Squartini and B. Schuller, Robust multi-stream keyword and non-linguistic vocalization detection for computationally intelligent virtual agents, *8th Int. Symp. on Neural Networks* (*ISNN 2011*), *Special Session "Computational Intelligence Algorithms for Advanced Human-Machine Interaction"* (IEEE Computational Intelligence Society, Springer Heidelberg, Guilin, China, 2011), pp. 496–505.
23. T. Drugman, J. Urbain and T. Dutoit, Assessment of audio features for automatic cough detection, *19th European Signal Processing Conference (EUSIPCO 11*) (Spain, 2011).
24. S. Petridis, M. Pantic and J. F. Cohn, Prediction-based classification for audiovisual discrimination between laugher and speech, *IEEE Int. Conf. on Automatic Face and Gesture Recognition and Workshops* (*FG 2011*) (2011), pp. 619–626.
25. T. Mikami, Y. Kojima, M. Yamamoto and M. Furukawa, Automatic classification of oral/nasal snoring sounds based on the acoustic properties, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP 2012*) (2012), pp. 609–612.
26. J. Zhang, B. Jiang, L. Lu and Q. Zhao, Audio segmentation system for sport games, *Int. Conf. on Electrical and Control Engineering* (*ICECE 2010*), pp. 505–508.
27. J. Huang, Y. Dong, J. Liu, C. Dong and H. Wang, Sports audio segmentation and classification, *IEEE Int. Conf. on Network Infrastructure and Digital Content* (*IC-NIDC 2009*) (November 6–8, 2009), pp. 379–383.
28. T. Perperis, T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. Tsekeridou, S. J. Perantonis and S. Theodoridis, Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies, *Expert Systems with Applications* **38**(11) (October 2011) 14102–14116.
29. T. Pohle, E. Pampalk and G. Widmer, Evaluation of frequently used audio features for classification of music into perceptual categories, in *Proc. of the Fourth Int. Workshop on Content-Based Multimedia Indexing* (*CBMI '05*) (2005).
30. M. Casey, General sound classification and similarity in MPEG-7, *Journal Organised Sound Archive* **6**(2) (August 2001) 153–164.
31. P. K. Ghosh, A. Tsiartas and S. Narayanan, Robust voice activity detection using long-term signal variability, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(3) (March 2011) 600–613.
32. C. V. Wright, L. Ballard, F. Monrose and G. M. Masson, Language identification of encrypted VoIP traffic: Alejandra y Roberto or Alice and Bob?, in *Proc. of 16th USENIX Security Symp. on USENIX Security Symposium* (*SS '07*) (2007).
33. B. Schuller, G. Rigoll and M. Lang, Speech emotion recognition combining acoustic feature and language information in hybrid suuport vector machine — Belief network architect-ture, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (2004). *Proceedings* (*ICASSP '04*), Vol. 1 (May 17–21, 2004), pp. 577–580.
34. I. Mporas, T. Ganchev, O. Kocsis and N. Fakotakis, Context-adaptive pre-processing scheme for robust speech recognition in fast-varying noise environment, *Signal Processing* **91**(8) (August 2011) 2101–2111.
35. F. Eyben, M. Wollmer and B. Schuller, openSMILE – The Munich versatile and fast open-source audio feature extractor, in *Proc. on the Int. Conf. on ACM Multimedia* (*MM 2010*) (ACM, Italy, 2010), pp. 1459–1462.
36. O. Lartillot and P. Toiviainen, A MATLAB toolbox for musical feature extraction from audio, in *Proc. of the 10th Int. Conf. on Digital Audio Effects* (*DAFx-07*) (Bordeaux, France, September 10–15, 2007).

37. M. Kos, M. Grasic, D. Vlaj and Z. Kacic, On-line speech/music segmentation for broadcast news domain, *16th Int. Conf. on Systems, Signal and Image Processing* (*IWSSIP 2009*) (June 18–20, 2009), pp. 1–4.
38. Y. Patsis and W. Verhelst, A speech/music/silence/garbage classifier for searching and indexing broadcast news material, *19th Int. Workshop on Database and Expert Systems Application* (*DEXA '08*) (September 1–5, 2008), pp. 585–589.
39. G. Richard, M. Ramona and S. Essid, Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP 2007*) (2007), pp. II-461–II-464.
40. V. Gupta, G. Boulianne, P. Kenny, P. Ouellet and P. Dumouchel, Speaker diarization of French broadcast news, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP 2008*) (March 31–April 4, 2008), pp. 4365–4368.
41. H.-Y. Lo, J.-C. Wang and H.-M. Wang, Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval, *IEEE Int. Conf. on Multimedia and Expo* (*ICME 2010*) (July 19–23, 2010), pp. 304–309.
42. M. Liu, C. Wan and L. Wang, Content-based audio classification and retrieval using a fuzzy logic system: Towards multimedia search engines, *Soft Computing* **6** (2002) 357–364.
43. G. Guo, H.-J. Zhang and S. Z. Li, Boosting for content-based audio classification and retrieval: An evaluation, *IEEE International Conference on Multimedia and Expo* (*ICME 2001*) (August 22–25, 2001), pp. 997–1000.
44. W. Pan, Y. Yao and Z. Liu, An unsupervised audio segmentation and classification approach, *Fourth Int. Conf. on Fuzzy Systems and Knowledge Discovery* (*FSKD 2007*), Vol. 3 (August 24–27, 2007), pp. 303–306.
45. H.-G. Kim, N. Moreau and T. Sikora, Audio classification based on MPEG-7 spectral basis representations, *IEEE Transactions on Circuits and Systems for Video Technology* **14**(5) (May 2004) 716–725.
46. F. Weninger, B. Schuller, M. Wollmer and G. Rigoll, Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory, *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP 2011*) (May 22–27, 2011), pp. 5840–5843.
47. C.-H. Wu and C.-H. Hsieh, Multiple change point audio segmentation and classification using an MDL-based Gaussian model, *IEEE Transactions on Audio, Speech and Language Processing* **14**(2) (March 2006) 647–657.
48. M. Markaki and Y. Stylianou, Discrimination of speech from nonspeech in broadcast news based on modulation frequency features, *Speech Communication* **53**(5) (2011) 726–735.
49. http://www.voanews.com/
50. http://www.itl.nist.gov/iad/mig/tests/lre/2009/.
51. The BBC sound effects library — original series. [Online], http://www.sound-ideas.com.
52. http://www.bbc.co.uk/podcasts/series/globalnews.
53. http://www.partnersinrhyme.com/soundfx/human.shtml.
54. http://soundbible.com/tags-laugh.html.
55. M. Slaney, Auditory Toolbox. Version 2. Technical Report #1998-010. Interval Research Corporation (1998).
56. K. Lee and M. Slaney, Automatic chord recognition from audio using an HMM with supervised learning, in *Proc. of the 1st ACM Workshop on Audio and Music Computing Multimedia* (*AMCMM '06*) (2006), pp. 11–20.
57. M. A. Bartsch and G. H. Wakefield, Audio thumbnailing of popular music using chroma-based representations, *IEEE Transactions on Multimedia* **7**(1) (2005) 96–104.
58. M. Robnik-Sikonja and I. Kononenko, An adaptation of relief for attribute estimation in regression, *14th Int. Conf. on Machine Learning* (*ICML '97*) (1997), pp. 296–304.

59. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, *Series B* **39**(1) (1977) 1–38.
60. I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. (Morgan Kaufmann, San Francisco, 2011).
61. T. M. Mitchell, *Machine Learning* (McGraw-Hill International Editions, 1997).
62. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation* **13**(3) (2001) 637–649.
63. D. Aha and D. Kibler, Instance-based learning algorithms, *Machine Learning* **6** (1991) 37–66.
64. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, CA, 1993).
65. L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci and A. Sarti, Scream and gunshot detection in noisy environments, in *Proc. of the 2007 IEEE Conf. on Advanced Video and Signal Based Surveillance* (*AVSS '07*) (2007), pp. 21–26.
66. D. L. Donoho, High-dimensional data analysis: The curses and blessings of dimensionality; Aide-Memoire of a lecture at AMS Conference on Math Challenges of the 21st Century (2000).
67. H. Mallinson and A. Gammerman, Imputation using support vector machines. Tech. rep., Royal Holloway, University of London (2003).
68. K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, When is "nearest neighbor" meaningful?, Database Theory — ICDT '99 (1999), pp. 217–235.
69. H. Hermansky, D. P. W. Ellis and S. Sharma, Connectionist feature extraction for conventional HMM systems, *Proc. ICASSP '00* (2000).
70. Q. Zhu, A. Stolcke, B. Y. Chen and N. Morgan, Using MLP features in SRI's conversational speech recognition system, *Proc. Interspeech '05* (2005).