MACHINE LISTENING TECHNIQUES AS A COMPLEMENT TO VIDEO IMAGE ANALYSIS IN FORENSICS

Romain Serizel, Victor Bisot, Slim Essid, Gaël Richard

LTCI, CNRS, Télécom ParisTech, Université Paris - Saclay, 75013, Paris, France

ABSTRACT

Video is now one of the major sources of information for forensics. However, video documents can be originating from various recording devices (CCTV, mobile devices...) with inconsistent quality and can sometimes be recorded in challenging light or motion conditions. Therefore, the amount of information that can be extracted relying solely on video image can vary to a great extent. Most of the videos however generally include audio recording as well. Machine listening can then become a valuable complement to video image analysis in challenging scenarios. In this paper, the authors present a brief overview of some machine listening techniques and their application to the analysis of video documents for forensics. The applicability of these techniques to forensics problems is then discussed in the light of machine listening system performances.

***Index Terms*—** Machine listening, source localisation, event detection, speaker identification, aoustic scene analysis, automatic speech recognition

1. INTRODUCTION

Video has recently become an increasingly important resource for forensics. Video captured by CCTV systems or video recorded from mobile devices (and possibly shared on multimedia platforms) can provide essential clues in solving criminal cases. For example when considering an investigation about a missing person, video documents can help to localise the missing person or a suspect, providing crucial information about their whereabouts. The analysis of videos linked with a missing person or her/his social network can also help to understand the conditions of the disappearance (was it a kidnapping, a runaway...) and largely influence the investigation.

However important they might be, video documents are generally recorded with various devices of unequal quality, in conditions that are often sub-optimal and with people or objects potentially masking the subject of interest. In such cases, the desired information might be difficult to retrieve based on visual content only. Yet most videos are recorded with audio and machine listening can be a valuable complement to video analysis in challenging scenarios.

Machine listening is a discipline at the interface of audio signal processing and machine learning that aims at automatically analysing and classifying audio recordings. Machine listening can include techniques relying purely on acoustic content such as acoustic scene classification (ASC), acoustic event detection (AED) or acoustic source localisation. It can also encompass to some extent speech analysis techniques such as speaker identification and automatic speech recognition (ASR). This paper intends to be a general introduction to machine listening as a set of complementary techniques for video analysis applications. The main targets are to briefly present a few machine listening techniques, to explain how they can extract information that is complementary to the information extracted with video analysis techniques and to discuss how the state-of-the-art approaches for machine listening can be relevant for forensics applications.

The paper is organised as follows. Section 2 introduces a few machine listening techniques: ASC, AED, acoustic source localisation, speaker identification and ASR, and present their potential applications to the analysis of video documents for forensics. Results recently obtained by the authors on ASC and speaker identification are recalled in Section 3 as an illustration of machine listening performance on selected tasks. The application of machine listening to video analysis in forensics in the light of the performance of state-of-the-art machine listening systems is discussed in Section 4 and conclusions are exposed in Section 5.

2. MACHINE LISTENING TECHNIQUES FOR VIDEO ANALYSIS

2.1. Acoustic scene classification

2.1.1. Description

ASC is the task of identifying in which acoustic environment a sequence was recorded based only on the audio signal (indoor, outdoor, street, train station, restaurant, office...). The interest for ASC has been increasing in the last few years and is becoming an important challenge in the machine listening community [1]. ASC has a variety of real life applications such as robotic navigation [2] or forensics [3]. Whilst many context aware devices only use visual information to adapt to their current location, complementary information can be given by analysing the surrounding audio environment. Major trends in ASC are to use various methods from speech recognition or event classification methods [4, 5, 6] or to use hand-crafted features designed to characterize acoustic environments [7, 8, 9]. We recently proposed to learn features for ASC in an unsupervised manner directly from time-frequency images and compared the performance of different approaches to learn these features [10].

2.1.2. Application on audio-visual recordings

Classifying video documents based on the type of scene where they were recorded is an essential step to perform video indexing. A properly indexed video database will allow faster information retrieval for forensics application on large scale data. Sometimes the visual information present in a video document may not be sufficient (recording angle is too narrow and centred, for example, on a person not on the surrounding) or might not be usable (poor light conditions, camera moving too fast) to perform accurate scene recognition. In such challenging cases, the audio recorded by the device might be a

This work was partly funded by the European Union under the FP7-LASIE project (grant 607480).

valuable complement to the video image analysis. ASC then allows to identify the type of scene where the video has been recorded and to index the video accordingly.

2.2. Acoustic event detection

2.2.1. Description

The target of AED is to detect specific events that occur in an audio recording and to identify the class of these events. The events are localised in time so the problem can be considered as twofold: detect the correct timing and the correct class. The AED has a variety of applications [11, 12] and can be strongly affected by the type of environment considered. For example indoor environments as offices [13] are usually considered as less challenging than outdoor real-life environments [14].

Event detection systems generally use standard acoustic features in a pattern recognition framework. Two different approaches exist. In a first approach, the events are detected independently of their class and are classified afterwards. The classifier then does not need to model time dependencies and can be for example a Gaussian mixture model (GMM) [15, 16]. In an alternative approach, detection and classification are performed jointly. The classifier then has to model time dependencies and a hidden Markov model (HMM) based classifier is commonly used [17, 18]. Recent approaches based on wavelets [19], bag of aural words [20] or Gabor filter-bank features [21] have allowed to reach higher performance.

2.2.2. Application on audio-visual recordings

An investigator looking for a video in a large dataset may want to retrieve information not only based on the type of scene where the video was recorded but also, at a finer granularity level, based on specific events that occurred during the recording. These events (such as whistle blowing, glass smashing, gunshot, cry...) are localised in time but they are often also localised in space. This means that the chances are high that they will not clearly appear in the video document. AED then allows to detect these events even if they do not appear visually or to confirm that an event that was only partially visible actually occurred.

In addition, the detection of specific events can help to confirm (or deny) the fact that a video was recorded in a particular scene. Some events are indeed representative of particular scenes for example train noise in all probability indicates the scene takes place in a train station or plates and cutlery noises indicate the scene is probably taking place in a restaurant. On the other hand, some events are unlikely to happen in particular scenes. AED can then help tracking anomalies to detect abnormal events (gunshots, crowd panic...) or to identify a recording scene where information has voluntary been concealed. This is the case, for example, when a kidnapper sends a ransom video recorded from inside a building but a church bell or a train passing nearby can be heard during the video. This type of information that is not present visually can help to localise the place where the video was recorded.

2.3. Acoustic source localisation

2.3.1. Description

The target in acoustic source localisation is to estimate the spatial position of one or several sources that are present in the acoustic scene recorded. There are two main classes of techniques in acoustic source localisation. The so-called "direct techniques" aim at estimating the direction of arrival (DOA) of a sound [22, 23, 24]. An intuitive way to estimate the DOA of a sound is to steer a beamformer at different potential directions and compare the results to the target signal to localise. In the second category of techniques the source localisation is estimated by proxy through the time difference of arrival (TDOA) of a sound at several microphones in an array with known geometry [25, 26]. Based on models about acoustic propagation, these TDOA then allow one to estimate the source localisation.

2.3.2. Application on audio-visual recordings

When several recordings from different spatial points in the same scene are available it is possible to consider localising audio sources spatially. This approach is used, for example, in gunfire locator systems. Acoustic source localisation can help to localise a person speaking in a scene. If the face of the person is not clearly visible on the video, it might be difficult to localise a speaker only from visual information. In this case machine listening provides a valuable complement to video analysis. Acoustic source localisation can also allow one to localise specific events and by proxy to refine the localisation of a person that was near a particular event when it happened but that is not present visually in the video document (or at least not in a way that allows for identification).

2.4. Speaker identification

2.4.1. Description

The main target of speaker identification is to assert whether or not the speaker of a test segment is known and if he/she is known, to find his/her identity [27]. Applications of speaker identification are numerous, among which speaker dependent automatic speech recognition and subject identification based on biometric information. The sentence pronounced by the subject is not necessarily known and the recordings can be of variable quality. The speaker identification then becomes a highly challenging problem.

Since their emergence almost five years ago, the I-vectors [28] have become the state-of-the-art approach for speaker identification [29]. A typical speaker identification system is composed of I-vector extraction, normalisation [30] and classification with probabilistic linear discriminant analysis (PLDA) [31]. Recent studies have shown that approaches such as nonnegative matrix factorisation (NMF) [32] can be successfully applied to retrieve speaker identity [33, 34]. Capitalising on these promising results, we have recently presented an approach deriving from the group-NMF [35] that intends to account for speaker variability and recording session variability by imposing constraints relatively to the speakers or recording sessions [36].

2.4.2. Application on audio-visual recordings

The identification of the persons present in a video is often a crucial aspect of video analysis for forensics. It can be useful to identify suspects, offenders, potential victims, hostages or missing persons. The problem is that the face of the persons involved in the recording cannot always be recognised visually. The poor quality of a video, the masking of the faces (intentionally or not) or simply the fact that the person of interest is the one recording the video are various obstacles to face recognition. When the persons involved in the video are speaking, speaker identification can help to confirm the identity of a person when the face recognition confidence is too low. The

joint analysis of video and audio can allow one to perform individual identification with higher accuracy. Speaker identification can also provide indication about the identity of a person whose face is concealed or who is not present visually in the video but whose speech has been recorded.

2.5. Automatic speech recognition

2.5.1. Description

The role of ASR is, given an audio recording, to automatically provide a transcription of what is said in the recording. Historically ASR systems used acoustic models based on HMM which appeared as a natural model for the sequential nature of speech. The emission probabilities of the HMM were then modelled from the acoustic feature vectors using GMM.

Recent advances in terms of training algorithms [37, 38] and computing power have lead to the generalisation of the use of deep neural networks (DNN) acoustic modelling and they are now the norm is ASR [39]. Consequently, the accuracy of the ASR systems have reached a point that makes them a credible technology to be used in mass market products [40, 41] and possibly in forensics applications. Two different paradigms compete for DNN-based acoustic modelling. In the first paradigm DNN are used to extract discriminative features that are to be used as input to the ASR [42, 43]. The idea behind the second paradigm is to use DNN to extract phonetic units from audio. The sequences of phonetic units that compose words are then modelled with HMM [39, 44].

2.5.2. Application on audio-video recordings

Speech is a structured and explicitly informative mean of communication. Speech in video can therefore carry a tremendous amount of information that can be difficult to recover relying only on visual content. Depending on the quality of the recording and the target application, ASR can allow to extract keywords from a recorded conversation or even in the best case to obtain its full transcription. Based on that information, it is possible to refine a summary of the video and to consider semantic indexing of video documents based on a set of selected keywords.

3. PERFORMANCE FOR SELECTED TASKS

To illustrate the performance that can be achieved on typical machine listening tasks, results recently obtained on ASC [10] and speaker identification [36] are reminded here. In both cases the multi-nomial logistic regression is used for classification and F1-score [45] is used as evaluation metric.

3.1. Acoustic scene classification

3.1.1. Evaluation corpus

The acoustic scene classification was evaluated on the LITIS Rouen data set [46]. It contains 25h of urban audio scenes recorded with a smart-phone, split into 3026 examples of 30s without overlap forming 19 different classes. Each class corresponds to a specific location such as *in a train station*, *in an air-plane* or *at the market*. The experiment protocol is the same as defined in Bisot *et al.* [10].

Method			F1-score
Previous state-of-the-art [48]			92.8%
Method	F1-score	Method	F1-score
PCA	89.9%	NMF	90.7%
Sparse PCA	90.0	Sparse NMF	**94.1%**
Kernel PCA	**95.6%**	Kernel NMF	84.1%
		Convolutive NMF	**94.5%**

Table 1. Weighted F1-scores obtained for a classification with multinomial logistic regression [10].

Features	I-vector	NMF	Group-NMF
F1-score	76.1%	70.7%	**80.2%**

Table 2. Weighted F1-scores obtained for a classification with multinomial logistic regression [36].

3.1.2. Results

In our recent paper [10] different popular matrix factorisation techniques are compared when used to perform unsupervised feature learning for acoustic scene classification. Experiments compare the use of extensions of the regular principal component analysis (PCA) [47] and NMF [32] such as sparsity, kernels and convolution. The classification scores are presented in Table 1 and show that these different variants of matrix factorization consistently improve the results over the standard approaches. The authors manage to outperform the previous state of the art results on the LITIS Rouen dataset [48] with Sparse NMF (94.1% F1-score), Kernel PCA (94.3% F1-score) and convolutive NMF (94.5% F1-score).

3.2. Speaker identification

3.2.1. Evaluation corpus

The speaker identification is evaluated on a subset of the ESTER corpus. ESTER is a corpus for automatic speech recognition composed of data recorded on broadcast radio [49]. The subset of ESTER used for evaluation is composed of 6 hours and 11 minutes of training data and 3 hours 40 minutes of test data both distributed among 95 speakers. The amount of training data per speaker ranges from 10 seconds to 6 minutes [36].

3.2.2. Results

In our recent paper [36] a state-of-the-art I-vector based speaker identification system is trained on the subset of ESTER with the LIUM speaker diarisation toolkit [50]. Its performance is compared to NMF-based systems: standard NMF and group-NMF. The systems parameters and evaluation protocol are similar to those described in Serizel *et al.* [36]. F1-scores are presented in Table 2. Variations in identification performance are validated using the McNemar test [51]. The first remark is that all systems perform reasonably well even if standard NMF is clearly behind the other approaches ($p < .001$). The group-NMF, by imposing constraints on both the speaker bases and the session bases, improves significantly the performance compared to the I-vector approach ($p < .01$).

4. DISCUSSION

Performances presented in this paper are obtained on corpora recorded in very specific conditions. Tests in real-life conditions

are difficult to set up and time consuming. It is therefore important to understand how these expirements can provide indications about the applicability of machine listening techniques in real-life video analysis for forensics. We consider here two main categories of recording devices: CCTV with audio and mobile devices (including smartphones, tablets, cameras...).

ASC achieves good performance on data recorded in realistic conditions, this would tend to indicate that ASC is mature to be applied to the analysis of documents recorded with mobile devices. A minor limitation however: recordings in the databases used to evaluate ASC usually include only the scene to be recognised (no perturbations) or at least the portions of the recodings when the scene is present alone are usually longer than what can be expected from real-life recordings. ASC will most likely become more challenging on short recordings or if another signal is dominating the recording (for example when performing ASC to classify the background environment of a recorded conversation).

AED can be related to ASC to some extent but is generally observed to be a more challenging task, especially when the number of classes increases or when events are overlapping in time. Classifiying very distinctive audio events (cries, gun shot...) is reliable as long as the events are in the foreground of the scene but the performance decreases drastically when the events are in the background of the recording [21, 52]. This latter scenario can occur for example when trying to detect events happening in the background of a recorded conversation. Having multiple recordings from the same scene could help to solve problems with events overlapping in time as the AED would then take advantage of the spatial localisation. The localisation itself can be quite robust when there is a control over the placement of the microphones (this could be the case when implanting new CCTV with audio).

Speaker identification has reached a high accuracy under very controlled conditions. It is therefore a credible techniques for documents recorded with close-up microphones. Indeed, most of the current techniques require to have at least a certain amount of clean (no noise) and dry (no reverberation) speech to achieve reliable identification. Performing recognition on distant speech with background noise, reverberation and possibly concurrent speakers can become really challenging. This can be seen as an obstacle to the analysis of video documents recorded with mobile devices but the research in speaker identification is moving towards more robust approaches that should allow this application [33]. However, speaker identification on video captured with CCTV with audio is by far more challenging and does not seem to be a viable option at the moment.

State-of-the-art ASR systems can achieve high performance on speech recorded with a close-up microphone and continuous speech transcription is then credible. Recent progress have been made in the domain of distant speech recognition [53] and ASR is a credible technique to analyse videos recorded with mobile devices. In more challenging scenarios (for example with lower signal-to-noise-ratio), considering keywords spotting instead of continuous speech transcription generally provides more robust systems that still allow semantic indexing of the videos. Yet, performing ASR on CCTV with audio does not seem to be a realistic option at the moment.

5. CONCLUSIONS

In this paper, we presented a brief overview of machine listening techniques and described their applications as a complement to video image analysis in forensics. In some challenging scenarios, when the video is degraded or when objects of interest are visually hidden, the machine listening techniques presented here can provide valuable information about the scene that was captured by the recording device. To illustrate the performance of state-of-the art systems in machine listening the performance recently obtained for selected machine listening tasks were presented. In the light of this performance, it appears that CCTV with audio could benefit from techniques such as AED and source localisation whereas ASC, speaker identification, ASR and to some extent AED could be helpful in the analysis of videos recorded from mobile devices. Machine listening could then be considered as an ideal companion to video processing. Ideally machine listening should even be used jointly with video image analysis for optimal performance.

6. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowel, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? scene recognition for mobile robots using audio features," in *Proc. of ICME*, 2006, pp. 885–888.

[3] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and mel-frequency cepstral coefficients," in *Proc. of ICDT*, 2010.

[4] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[6] X. Valero and F. Alías, "Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

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

[8] R. Mogi and H. Kasaii, "Noise-robust environmental sound classification method based on combination of ica and mp features," *Artificial Intelligence Research*, vol. 2, no. 1, pp. p107, 2012.

[9] S. Deng, J. Han, C. Zhang, T. Zheng, and G. Zheng, "Robust minimum statistics project coefficients feature for acoustic environment recognition," in *Proc. of ICASSP*, 2014, pp. 8232–8236.

[10] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorisation for unsupervised feature learning," in *Accepted for publication in Proc. of ICASSP*, 2016.

[11] P. Volgyesi, G. Balogh, A. Nadas, C. B. Nash, and A. Ledeczi, "Shooter localization and weapon classification with soldier-wearable networked sensors," *Mobisys*, p. 113, 2007.

[12] Q. Huang and S. Cox, "Hierarchical language modeling for audio events detection in a sports game," *Proc. of ICASSP*, pp. 2286–2289, 2010.

[13] D. Giannoulis and E. Benetos, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," *Proc. of WASPAA*, 2013.

[14] T. Heittola and A. Mesaros, "Sound event detection in multisource environments using source separation," *Workshop on machine listening in Multisource Environments*, pp. 35–40, 2011.

[15] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," *Proc. of ICME*, vol. 2005, pp. 1306–1309, 2005.

[16] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," *Proc. of EUSIPCO*, pp. 1216–1220, 2007.

[17] J. T. Geiger, M. A. Lakhal, B. Schuller, and G. Rigoll, "Learning new acoustic events in an HMM-based system using MAP adaptation," *Proc. of Interspeech*, pp. 293–296, 2011.

[18] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," *Proc. of ICASSP*, 2009.

[19] A. Chacón-Rodríguez, P. Julián, L. Castro, P. Alvarado, and N. Hernández, "Evaluation of gunshot detection algorithms," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 2, pp. 363–373, 2011.

[20] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," *Proc. of AVSS*, pp. 81–86, 2013.

[21] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in *Proc. of EUSIPCO*, Aug 2015, pp. 714–718.

[22] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[23] N. Epain and C. T. Jin, "Super-resolution sound field imaging with sub-space pre-processing," *Proc. of ICASSP*, 2013.

[24] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields.," *The Journal of the Acoustical Society of America*, vol. 112, no. 5 Pt 1, pp. 1980–1991, 2002.

[25] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," *Proc. of ICASSP*, vol. 3, 2004.

[26] C. M. Zannini, C. M. aZannini, R. Parisi, and A. Uncini, "Improved TDOA disambiguation techniques for sound source localization in reverberant environments," *Proc. of ISCAS*, pp. 2666–2669, 2010.

[27] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Advances in Signal Processing*, vol. 4, pp. 430–451, 2004.

[28] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[29] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.

[30] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition.," in *Proc. of Interspeech*, 2011, pp. 485–488.

[31] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of ICCV*, 2007, pp. 1–8.

[32] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[33] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Noise Robust Speaker Recognition with Convolutive Sparse Coding," in *Proc. of Interspeech*, 2015.

[34] N. Seichepine, S. Essid, C. Fevotte, and O. Cappe, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[35] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proc. of AISTATS*, 2009, pp. 320–327.

[36] R. Serizel, S. Essid, and G. Richard, "Group nonnegative matrix factorisation with speaker and session variability compensation for speaker identification," in *Accepted for publication in Proc. of ICASSP*, 2016.

[37] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[38] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[39] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[40] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[41] G.E. Dahl, Dong Y., Li D., and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[42] H. Hermansky, D. PW Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. of ICASSP*, 2000, vol. 3, pp. 1635–1638.

[43] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proc. of ICASSP*, 2007, vol. 4, pp. IV–757.

[44] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, vol. 247, Springer, 1994.

[45] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, 1979.

[46] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[47] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1, pp. 37–52, 1987.

[48] V. Bisot, S. Essid, and G. Richard, "Hog and subband power distribution image features for acoustic scene classification," in *Proc. of EUSIPCO*, 2015.

[49] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri, "ESTER, une campagne d'évaluation des systemes d'indexation automatique d'émissions radiophoniques en français.," in *Proc. of Journées d'Etude sur la Parole*, 2004.

[50] M. Rouvier, G. Dupuy, P. Gay, and E. Khoury, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. of Interspeech*, 2013.

[51] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[52] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. of ICASSP*, 2015.

[53] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proc. of ASRU*, 2015.