TWO-CLASS VERIFIER FRAMEWORK FOR AUDIO INDEXING

V. Ramasubramanian[1], *S. Thiyagarajan*[1], *G. Pradnya*[1*], *Heiko Claussen*[2], *Justinian Rosca*[2]

[1] Siemens Technology and Services, Research and Technology Center, Bangalore, India
[2] Siemens Corporation, Corporate Technology, Princeton, USA

ABSTRACT

We address the problem of audio indexing for a class of special-case scenarios, where it is required to index an audio stream into 2 classes, namely, a target class and a background class, as arising, say in, audio surveillance and machine diagnostics. With the emphasis on dealing with limited training exemplars defining the target class in these scenarios, we propose a 2-class 'audio verification' framework, where the target and background classes are modeled by GMMs and the indexing is done via a sliding window based detection. We characterize the performance of the system in terms of ROCs, EERs and visual detection plots for a set of 2 target classes and 4 background classes from a surveillance audio database and show the viability of such a system in practical applications. We highlight the robustness of the system to high levels of background-class using visual detection plots of continuous audio streams at SNRs ranging from 30 dB down to -20 dB.

Index Terms— Audio indexing, 2-class verification, audio verification, surveillance audio, machine diagnostics

1. INTRODUCTION

Audio indexing, in the conventional sense, involves segmenting and labeling (the segments) of an input audio stream into one or more pre-defined audio classes. It is generally assumed that the audio stream is made of a sequence of non-overlapping audio sounds occurring in some unspecified order and of varying durations each. The audio sounds are also assumed to belong to a set of N 'vocabulary' audio classes. In a variation of this scenario, it is possible to consider the sequence of audio sounds (drawn from the pre-defined vocabulary of audio classes) to occur interspersed in a background audio (some other audio classes), thereby giving rise to a segmentation and labeling problem in terms of the vocabulary audio-classes and a background class making up the (possibly larger proportion of the) audio stream.

For arbitrary N ($N > 1$), the above problem is best handled in a manner similar to watch-list based detection frameworks, i.e., in an open-set identification framework (or the multi-target identification framework) [1], [2], [3]. In contrast, in this paper, we deal with a refinement of the above scenario as a special case variant of the above definition with $N = 1$, i.e., the audio vocabulary has only one audio class (called the 'target' class of interest) to be detected in an input audio stream predominantly made of a background class. Thus, the input audio stream is essentially a background class in which one specific 'target' audio class occurs at unspecified times and of varying durations from instance to instance, theretby reducing the problem to a 2-class indexer in terms of alternating target and background labels. Fig. 1 illustrates the general set-up of a 2-class verifier system. In the following, we present two scenarios where this system in used.

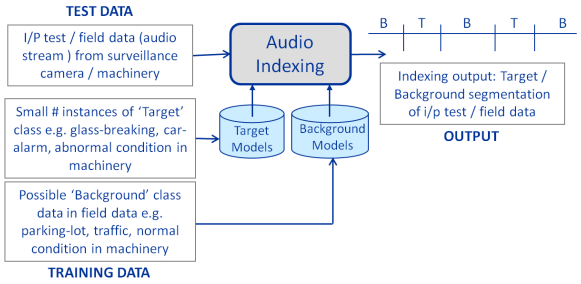

Fig. 1. *Generic audio indexing for 2 class special case scenarios.*

Scenario 1: In a typical security and surveillance context, a surveillance camera provides a video and audio stream of the place under its purview. A large number of events could take place, e.g., for a camera in a parking-lot, the various events would be car pass by, car braking, car horn, footsteps, babble, car door open/shut sounds, etc.. Here, it is of interest to detect a critical event, say glass-breaking, as when an intruder attempts to break-in to a car by breaking a car door glass or car-alarm, even while considering all other sounds as not of interest, i.e., possibly making up the 'background' class, made of multiple heterogeneous sounds. Such an audio indexing system can operate both on-line and off-line, generating actionable intelligence of different nature. For example, a) on-line, when it is indexing a live audio feed from a camera in order to detect a live break-in, to further trigger an alarm, operator alert, camera feed recording or other actions, or b) off-line, when it is required to index a stored camera feed - possibly a day long recording at the end of the day - so as to generate an indexed data, with the index information, i.e., segmentation and labeling, being stored as meta data along with the raw camera audio and video feed for future retrieval as in forensic search requirements.

Scenario 2: Industrial machinery (e.g. turbines, wind-mills etc.) generate considerable number of acoustic signatures each of which carries significant information on the health of the machinery (as a whole or specific parts generating the said acoustic signature). The machinery is expected to be in normal conditions much of the time, with 'abnormal conditions' (deviant states or faulty conditions) occurring intermittently or rarely. Therefore, an audio stream generated via an appropriate sensor can be considered to be largely made of a background class representing the 'normal condition' while being interspersed with occasional occurrences of the acoustic signatures of the 'abnormal conditions'. The audio corresponding to the abnormal conditions can be considered to define the 'target' class (of which there could be several types and instances, i.e. with high acoustic variability). The audio representing the normal condition then defines the 'background' class (typically characterized by no or

*On internship at Siemens Technology and Services, Research and Technology Center, Bangalore

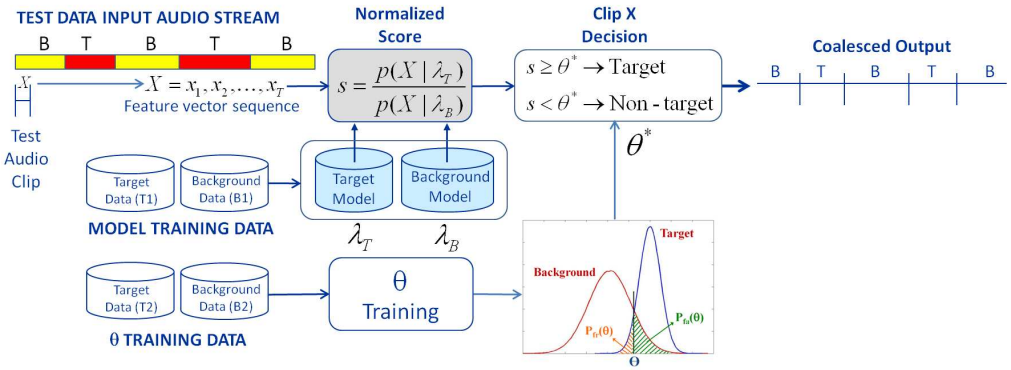

Fig. 2. *Schematic of the 2-class audio verification framework.*

minimal acoustic variability). An audio indexing system plays the following role here: it is required to index an input audio stream in terms of normal and abnormal classes, in the form of a segmentation and labeling, essentially detecting the occurrences of the abnormal conditions (target classes), with such target detections forming actionable intelligence for triggering an operator alert or other machine diagnostic services to be brought in.

A further consideration in the above indexing scenarios is the availability of training exemplars of the target classes. It can be noted that in either of the two scenarios above, the target class data's availibility is limited with respect to being able to build robust models (i.e., during a training phase) required for further classification and detection during test conditions. For instance, in the case of the surveillance scenario, for a target class as 'glass-breaking' (car-glass or retail windows), it is fairly impractical to expect large amounts of training data due to the difficulty in acquiring such data by actually breaking glasses; note for instance, the arduous data collection protocol involved in the *Glass Break Sound Pack*^{TM} development of a commercial solution [4]. Likewise, in the second scenario of machine diagnostics, by definition, the abnormal conditions making up the target class occur infrequently (or even not at all), thereby making training data availability scarce.

In this paper, we propose a 2-class verifier (or simply an 'audio verification') framework wherein the target class models are built as GMMs from limited training data and used alongside background class models (trained from possibly larger amount of training data) in a verification framework. A short sliding window (clip) of the input audio is verified as being the target class or not (i.e., background class). Successive clips with same decision labels are coalesced into a larger segment bearing the same label to yield the desired indexing in terms of alternating target and background classes.

2. TWO-CLASS VERIFICATION FRAMEWORK

Fig. 2 shows the schematic of the 2-class verification system proposed here for segmentation and labeling of the 2-class audio data. This system is similar to a speaker verification system, involving the following main steps:

Step 1 - Model training: The target and background models - Gaussian mixture models (GMMs) λ_T, λ_B - are first trained from the target and background 'model-training' data (T_1, B_1) using EM algorithm [5].

Step 2 - θ-training: Obtaining the optimal decision threshold θ^* that defines the equal-error-rate operating point for the target and background 'θ-training' data sets (T_2, B_2). This θ-training data is used to derive the normalized target and background score histograms from which the probability of false-alarms $p_{fa}(\theta)$ and the probability of false-rejections $p_{fr}(\theta)$ are obtained for various θ values. This yields the receiver operating characteristic (ROC) curve of $p_{fr}(\theta)$ vs $p_{fa}(\theta)$ and the equal-error-rate (EER) point when $p_{fa}(\theta^*) = p_{fr}(\theta^*) = \text{EER}$. The optimal θ^* corresponding to EER on the θ-training data set is used further for testing on unseen test audio stream as shown in Fig. 2.

Step 3 - Testing: Test data is considered in the form of a continuous audio stream made of alternate target (T) and background (B) classes (as shown in alternating red and yellow strips respectively in Fig. 2). The 2-class verification decision is obtained on sliding short clips of data $X_1, X_2, \ldots, X_i, \ldots, X_M$, i.e., obtain target (T) / background (B) verification decision for clips $X_i, i = 1, \ldots, M$, as follows:

1. For each clip X, obtain feature vector sequence $X = (x_1, x_2, \ldots, x_t, \ldots, x_T)$. Here, x_t is a mel frequency cepstral coefficients (MFCC) vector of dimension $d = 12$, obtained on a frame-size of 20ms, yielding $T = 10$ feature-vectors/clip for clip duration of 200ms.
2. Verify each clip X as target (T) or background (B) class:
   - Compute normalized score $s = p(X|\lambda_T)/p(X|\lambda_B)$, where $p(X|\lambda_T)$ and $p(X|\lambda_B)$ are the likelihoods of the vectors $\{x_t\}_{t=1,\ldots,T}$ in the test clip X to the GMMs λ_T and λ_B respectively.
   - $s \geq \theta^* \rightarrow X$ is target (T)
   - $s < \theta^* \rightarrow X$ is non-target (background B)
   - Decision label $L(X) = \text{T or B}$

Above verification for $(X_1, X_2, \ldots, X_i, \ldots, X_M)$ yields clip-level label sequence $L(X_1), L(X_2), \ldots, L(X_i), \ldots, L(X_M)$. Subsequent to clip-level decisions as above, contiguous clips with same decision are coalesced to yield segment-level labels; i.e., if $L(X_i) = L(X_{i+1}) = \ldots = L(X_{i+j}) = \text{T}$, then $L(X_i, \ldots, X_{i+j}) = \text{T}$, to yield the final target / background segmentation as shown as $(\ldots \text{BTBTB} \ldots)$ in the figure as the final output of the verifier system. The actual system performance on unseen test data is given by $(p_{fa}(\theta^*), p_{fr}(\theta^*))$, which reflects the (false-alarm, false-rejection) performance when the system uses the optimal θ^* for verifying unseen test data, as in field conditions.

3. EXPERIMENTS AND RESULTS

In this work, noting the practical utility of glass-breaking and car-alarm sounds in a surveillance scenario, we have used 2 target classes, namely, glass-breaking and car-alarm and 4 background classes, namely, market, babble, footsteps and public announcement (PA). We obtained training (separate data sets for model and θ-training) and test data for these 6 classes from the databases BBC Sound Effects Library, Series 1000 and Series 6000 [6].

Here, we present results characterizing the overall performance of the system in terms of i) EERs and ROC curves of the system as obtained during θ-training on data sets (T_2, B_2), ii) $(p_{fa}(\theta^*), p_{fr}(\theta^*))$ obtained on test data, and iii) decoding plots on test data, comparing the coalesced final output of the verifier with the ground truth in the form of time synchronous double-color strips for select target, background class combinations and at different target-to-background SNRs.

Table 1. *2-class verifier performance: EERs during θ-training for the 2 target classes in 4 different background classes.*

Target Class	Background class			
	Babble	Footsteps	Market	PA
Glass Breaking	9.85	19.83	3.28	0
Car Alarm	3.35	0.91	6.99	0.51

Table 1 shows the EER of the θ-training stage for the 2 target classes, each in combination with 4 different background classes. The EER is obtained as $(p_{fa}(\theta^*)+p_{fr}(\theta^*))/2)$, i.e., the mean of the false-alarm and false-rejection probabilities (in %) at the optimal θ^*. We also show in Fig. 3 the corresponding ROC curves obtained for these 2 target classes with each of the 4 background classes during θ-training. The EERs marked in the ROCs in Fig. 3 (a) and (b) are as shown in Table 1.

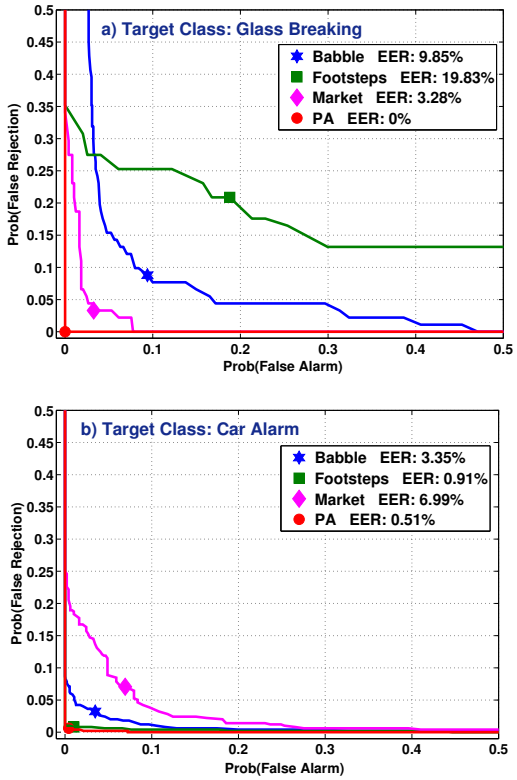

Fig. 3. *ROC plots of the 2-class verifier for 2 target classes a) glass-breaking and b) car-alarm in 4 different background classes, obtained during θ-training.*

A range of performance can be noted, primarily from the standpoint of how distinct the target class is with respect to the background class in any given target-background combination: i) Glass-breaking performs best in 'PA' background; it also performs surprisingly well in 'Market' background despite the heterogenous mix of sounds that the market class is made of, ii) Car-alarm performs with a consistently low EER across all background classes, owing to its unique signature. These observations are to be qualified further by noting that the GMM modeling of the classes is static in nature and does not capture the underlying temporal dynamics and signatures of the classes involved. Thereby the matching scores obtained, at best, capture only the spectral characteristics without being able to discriminate the classes using the temporal evolution of the sound spectra, as will template models or hidden Markov models (HMM); this was noted in our earlier work [20], where HMMs outperformed vector quantization (VQ) and GMM based audio decoding.

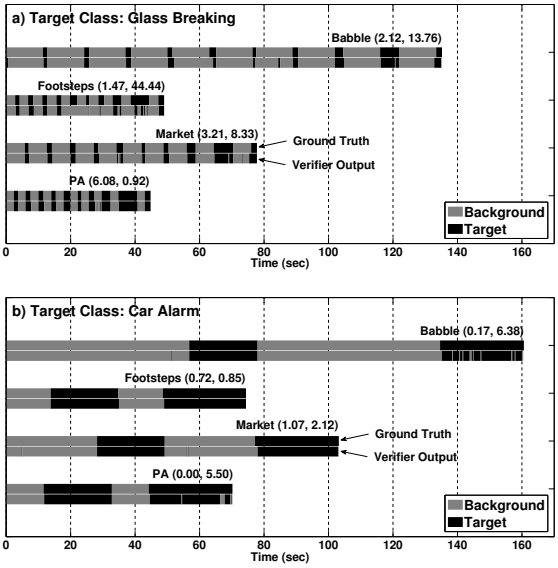

Fig. 4. *Continuous test audio stream decoding performance of the 2-class verifier for a) 'glass-breaking' and b) 'car-alarm' in 4 different background classes. In each figure, top strip is the ground truth of test audio stream and adjoining bottom strip is the decoding output.*

Fig. 4 shows the decoding of continuous test audio stream as a strip-plot (with ground truth as the top strip and the 2-class verifier output as the adjoining bottom strip) using the θ^* derived from the ROC plots in Fig. 3. The following can be noted with reference to the decoding strip-plot in Fig. 4:

i) The target class segments are detected quite accurately, and can be considered acceptable for a practically useful system. For glass-breaking during footsteps, 2 of the 10 target instances are missed contributing to the high false-rejections. Apart from this, the system detects all the 'glass-breaking' and 'car-alarm' segments accurately for all the background classes.

ii) Noted alongside the 8 strip-plots in Fig. 4 are the $(p_{fa}(\theta^*), p_{fr}(\theta^*))$ [%] for each of the 8 cases. A prominent pattern across most of the target-background combination is that $(p_{fa}(\theta^*), p_{fr}(\theta^*))$ is skewed (i.e., lower $p_{fa}(\theta^*)$ and higher $p_{fr}(\theta^*)$) than the corresponding training EER. While the detections are accurate, a closer observation of the strip-plots reveals that this skewing primarily arises from a number of misclassified 'test clips' (thin vertical lines with mismatching color) - short sporadic instances or extensions of target/background segments. Such a behavior therefore does not detract from the good primary detections as is visibly seen in Fig. 4 which translate into correct actionable intelligence in a practical system. However, this also leads to the conclu-

sion that the system performance can be improved significantly by merely targeting and correcting for this skewed $(p_{fa}(\theta^*), p_{fr}(\theta^*))$ by any of several means: a) use of longer clips X in scoring, b) use of variable length clips after model-free change detection [7], c) use of duration constraints for smoothing and integrating the individual clip decisions into the longer 'coalesced' decision.

iii) Another important causative factor of the skewed $(p_{fa}(\theta^*), p_{fr}(\theta^*))$ is the training-test mismatch, where we have observed the 'test' data to have high variability (with respect to the data used in GMM model training and θ-training) with resultant low scores (low GMM likelihoods) of the test clips with the GMMs and a corresponding left-ward shift in the target-score and background-score histograms with respect to the histograms obtained during θ-training. Consequently, a good proportion of the 'test' background clips have scores less than the operating threshold θ^* and thereby a lowered $p_{fa}(\theta^*)$. Likewise, a higher $p_{fr}(\theta^*)$ results from the lower scores (lower GMM likelihoods) of the 'target' test clips, i.e., a corresponding left-ward shift of the target score histogram and consequent increase in the number of test clips with scores lesser than θ^* leads to more target clips being falsely-rejected. This becomes evident in the relatively larger proportion of very short lines of black (target clips/segments) being classified as grey (background). As this clearly points to the training-test mismatch as a major contributing factor, this is best addressed by 'model adaptation'. That is, an on-line unsupervised adaptation, as is common in on-line speech recognition systems [8], [9] using adaptation methods such as MAP, MLLR etc. [10], [11], to adapt the model (GMM) parameters and match them to the new test data continuosly.

Next, we examine the robustness of the system to varying levels of background class, by treating it as additive noise with respect to the foreground target class. For this, we created continuous stream sound mixtures of 10 different target instances (glass-breaking) in a long backgrouund audio (market), at SNRs ranging from 30 dB to -20 dB in steps of 5 dB. Fig. 5 shows the decoding results of the 2-class verifier for such continuous test audio streams at various SNRs. All 10 target segments are 'detected' from 30 dB down to 10 dB, 7 segments at -5 dB SNR, 6 segments at -10 dB SNR and 3 segments at -20 dB SNR. This performance is satisfactory from a practical standpoint, where ambient conditions with -5 dB SNR are a good upper bound on the background level. However, SNRs above 10 dB are preferable for applications that require close to 100% detection accuracy such as triggering of on-line operator alerts or reliable off-line forensic search.

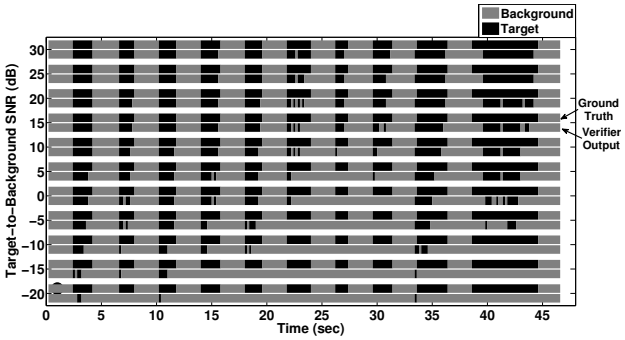

Fig. 5. *Decoding performance of 2-class verifier for SNR data: target class as 'signal' and background class as 'noise' at SNRs ranging from 30 dB down to -20 dB.*

4. RELATION TO PRIOR WORK

The 2-class indexing proposed here can be viewed as a constrained special case of a multi-class audio indexing as has been studied in earlier work. Specifically, much of the earlier work has focussed on 'isolated instance' multi-class classification using various types of modeling and classification, such as K-nearest-neighbor (K-nn) classification [12], [13], decision tree clsasification [13], quadratic Gaussian classifier [13], GMM [12], [20], templates [19] and HMM [14], [15], [16], [17], [18], [20]. Specifically, in [19] and [20], we proposed audio 'decoding' solutions using template based modeling and HMM based modeling, wherein an incoming audio stream made of multiple audio classes can be segmented and labeled from the given vocabulary of audio classes using either the one-pass dynamic programming decoding algorithm or the Viterbi decoding algorithm.

Our approach to audio indexing reduces the problem of multi-class classification and indexing to a two-class segmentation and labeling into a specific target class and background representing a non-target class. More specifically, the proposed two-class verifier aims to detect a target class of interest occurring in a heterogeneous background made of possibly many types of atomic sounds. For example, each of market, traffic or parking-lot audio superimposes a variety of atomic classes such as car-pass by, car-horn, car-screech, car-braking, door open/shut, babble, footsteps etc. Rather than attempting to model such a heterogeneous background class in terms of the individual audio classes from a multi-class vocabulary, we shift the emphasis towards an efficient two-class verification system that essentially makes a target/non-target decision for every short clip of input audio, with the non-target decision being equivalent to the input clip being classified as background class. In other words, the focus shifts to how efficiently the background class can be modeled. This is very much akin to the speaker verification problem of deciding on an input speech as from a claimant speaker (target) or an impostor (non-target), with the emphasis being laid on how effectively to model the impostor (non-target) population by using universal background models (UBMs) or cohorts [21], [22], [23].

In this context, it is appropriate to note that an alternate approach to this 2-class indexing could be to use 1-class SVM [24], whose 'positive data set' is defined to comprise the background class (whose training data is adequately available) and which therefore can detect the target class - even with no training data - as an outlier-detection (or also termed novelty-detection) problem. However, given the availability of both limited target data and adequate background data as in the scenarios considered here, the 2-class verification framework proposed and studied here can be expected to be more appropriate, efficient and offer better performance.

5. CONCLUSIONS

We have proposed a 2-class audio-verification system for audio indexing of special case scenarios such as arising in surveillance and machine diagnostics settings. Here it is of interest to detect a target class occurrance amidst a background which can potentially be considered a single class, though possibly made of a large number of unknown and heterogenous classes, but distinct from the foreground target class of primary interest. We have studied the performance of the system for 2 target and 4 background classes in a surveillance setting and characterized the performance of the system in terms of ROCs, EERs and the actual (false-alarm, false-rejection) probabilities on unseen test data. We have shown that the system offers high performance with practically useful detection ability down to 10 dB SNR and acceptable performance even for SNR down to -5 dB representing very high background noise levels.

6. REFERENCES

[1] E. Singer and D. Reynolds. Analysis of multi-target detection for speaker and language recognition. *Proc. Odyssey 2004 The Speaker and Language Recognition Workshop*, Toledo, 2004.

[2] Y. Zigel and M. Wasserblat. How to deal with multiple-targets in speaker identification systems?. *Proc. Odyssey 2006 The Speaker and Language Recognition Workshop*, San Juan, 2006.

[3] V. Ramasubramanian. Speaker Spotting: Automatic Telephony Surveillance for Homeland Security. Ch. 15, pp. 427-468, Forensic Speaker Recognition, A. Neustein, H. A. Patil (eds.), Springer Science+Business Media, LLC 2012.

[4] *Glass Break Sound Pack*^{TM} from Audio Analytics. http://www.audioanalytic.com/en/component/content/article/12-content/3-sound-pack-development

[5] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, Jan 1995.

[6] http://www.sound-ideas.com/bbc.html, http://www.sound-ideas.com/1000.html, http://www.sound-ideas.com/6000.html

[7] S. Chen, P. Gopalakrishnan. Speaker, environment, and channel change detection and clustering via the Bayesian information criterion. In: Proc. DARPA speech recognition workshop, pp. 127-132, 1998.

[8] X. Huang, A. Acero, H. W. Hon. Spoken Language Processing: A guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[9] J. Droppo and A. Aceoro. Environmental robustness. Ch 33, In: J. Benesty, M. M. Sondhi, Y. Huang (eds), Handbook of Speech Processing, Springer-Verlag Berlin Heidelberg, pp. 653-679, 2008.

[10] Chin-Hui Lee and Jean-Luc Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In *Proc. ICASSP' 93*, pp. 558-561, Minneapolis, Minnesota, 1993.

[11] C. J. Legetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech and Language*, vol. 9, pp. 1-17, 1995.

[12] V. Peltonen et al. Computational auditory scene recognition. *Proc. ICASSP '02*, Orlando, Florida, 2002.

[13] K. El-Maleh, A. Samouelian and P. Kabal. Frame-level noise classification in mobile environments. *Proc. ICASSP '99*, 1999.

[14] P. Gaunard et al. Automatic classification of environmental noise events by hidden Markov models. *Proc. ICASSP '98*, 1998.

[15] L. Ma, D. J. Smith and B. P. Milner. Context awareness using environmental noise classification. *Proc. Eurospeech '03*, pp. 2237-2240, Geneva, Switzerland, 2003.

[16] L. Ma, B. Milner and D. Smith. Acoustic environment classification. ACM Transactions on Speech and Language Processing, vol. 3, no. 2, pp. 1-22, July 2006.

[17] P. Nordqvist and A. Leijon. An efficient robust sound classification algorithm for hearing aids. Journal of Acoustical Society of America, vol. 115, no. 6, pp. 1-9, June 2004.

[18] Z. Liu, J. Huang and Y. Wang. Classification of TV programs based on audio information using hidden Markov model. *IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 27-32, Redondo Beach, CA, Dec 1998.

[19] Srikanth Cherla and V. Ramasubramanian. Audio analytics by template modeling and 1-pass DP based decoding. Proc. Interspeech 2010, pp. 2230-2233, Chiba, Japan, Sep 2010.

[20] V. Ramasubramanian, R. Karthik, S. Thiyagarajan and Srikanth Cherla. Continuous audio analytics by HMM and Viterbi decoding. Proc. ICASSP' 11, pp. 2396-2399, Prague, Czech Republic, May 2011.

[21] D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, vol. 10, no. 1, pp. 19-41, 2000.

[22] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, D. A. Reynolds. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, 4:430-451, 2004.

[23] D. A. Reynolds, W. M. Campbell. Text-independent speaker recognition. In: J. Benesty, M. M. Sondhi, Y. Huang (eds), Handbook of Speech Processing, Springer-Verlag Berlin Heidelberg, pp. 763-781, 2008.

[24] D. M. J. Tax. One-class classification: Concept-learning in the absence of counter-examples. Ph.D. thesis, Delft University of Technology (2001).