Pattern Recognition Letters 31 (2010) 1543–1551

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Real-world acoustic event detection

Xiaodan Zhuang *, Xi Zhou, Mark A. Hasegawa-Johnson, Thomas S. Huang

Beckman Institute of Advanced Science and Technology, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Available online 19 February 2010

Keywords:
Acoustic Event Detection
Feature selection
Hidden markov model
Artificial neural network
Tandem model
Gaussian mixture model supervector

ABSTRACT

Acoustic Event Detection (AED) aims to identify both timestamps and types of events in an audio stream. This becomes very challenging when going beyond restricted highlight events and well controlled recordings. We propose extracting discriminative features for AED using a boosting approach, which outperform classical speech perceptual features, such as Mel-frequency Cepstral Coefficients and log frequency filterbank parameters. We propose leveraging statistical models better fitting the task. First, a tandem connectionist-HMM approach combines the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained using the minimum cross entropy criterion. Second, an SVM–GMM-supervector approach uses noise-adaptive kernels better approximating the KL divergence between feature distributions in different audio segments. Experiments on the CLEAR 2007 AED Evaluation set-up demonstrate that the presented features and models lead to over 45% relative performance improvement, and also outperform the best system in the CLEAR AED Evaluation, on detection of twelve general acoustic events in a real seminar environment.

1. Introduction

Much research in audio content analysis has typically addressed the problem of segregating a few audio sources (Brown and Cooke, 1994; Ellis, 1996) or segmenting an audio stream into a small number of acoustically compact categories (Pinquier, 2002; Zhang and Kuo, 2001). Acoustic Event Detection (AED) aims to detect specified acoustic events such as gunshots (Clavel et al., 2005), explosions (Naphade, 2001; Cui et al., 2003a), speech/music transitions (Pinquier, 2002), cough events (Smith et al., 2006), or audience cheering at a sports event (Baillie and Jose, 2003). The existence and timestamps of many non-speech sounds, i.e. (non-speech) acoustic events, reveal human and social activities. Such information is very helpful in applications such as surveillance, multimedia information retrieval and intelligent conference rooms.

While most of the work in acoustic event detection focuses on a few highlight acoustic events, the 2007 AED Evaluation sponsored by the project "Classification of Events, Activities and Relationships (CLEAR)" (Temko et al., 2006; Temko, 2007) was performed on a continuous audio database recorded in real seminars (Temko and Nadeu, 2005). Systems attempted to identify both the temporal boundaries and labels of twelve acoustic events (door slam, paper wrapping/rustling, foot steps, knocking, chair moving, phone ringing, spoon/cup jingle, key jingle, keyboard typing, applause, cough, and laughter). Instead of being exclusively highlight events, many of the acoustic events in the CLEAR Evaluations were either subtle (low SNR, e.g. steps, paper wrapping/rustling, and keyboard typing), or/and overlapping with speech, making the task particularly challenging. The real environment factor added to the variation of the events as well as the difficulty of segmenting the audio stream. Although different system architectures and feature sets have been explored (Temko et al., 2006; Temko, 2007), even the top rated AED system (around 30% accuracy) left much space for improvement (Zhou et al., 2007). By contrast, classification of performed isolated events in silent rooms saw very good performance achieved by some of the same research teams (Temko et al., 2006). The evaluation highlighted the challenges in the detection of a large set of ordinary acoustic events in a real world environment.

To tackle AED in such a realistic setting, we believe further improvement is possible with features and statistical models better fitting the task, drawing lessons from the CLEAR 2007 AED Evaluation. A small part of this work was previously reported (Zhou et al., 2007; Zhuang et al., 2008).

Analysis of the spectral structure of acoustic events and design of a suitable feature set are important for AED. Various audio perceptual features have been proposed for different analysis tasks (Brown and Cooke, 1994; Scheirer, 1999; Cui et al., 2003b). In the recent CLEAR Evaluations for AED, the most popular features are speech perception features (Temko et al., 2006; Atrey et al., 2006), such as Mel-Frequency Cepstral Coefficients (MFCC) and log frequency filter bank parameters, which have been proven to represent speech spectral structure well. However, these features

* Corresponding author. Tel.: +1 217 898 6732; fax: +1 217 244 9233.
E-mail addresses: xiaodan.zhuang@gmail.com (X. Zhuang).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.02.005

are not necessarily suitable for AED for the following reasons. First, limited work has been done in studying the spectral structure of acoustic events. The speech features designed according to the spectral structure of speech might be far from optimal for AED. Second, the Signal-to-Noise Ratio (SNR) is low for AED especially when the overlapping speech can be seen as noise.

In this study we propose a new front-end feature analysis and selection approach for AED. Considering the varying discriminative capabilities of each feature component for the AED task, we propose a boosting approach to construct a discriminative feature set from a large feature pool.

AED in real seminars differs from classification of isolated events in a silent environment, calling for different statistical models. While SVMs were shown to be optimal for the latter (Schòlkopf and Smola, 2002), the former saw most leading CLEAR participants using dynamic Bayesian networks (Temko et al., 2006; Temko, 2007), in particular, hidden markov models (HMMs). HMMs owe their success to the Viterbi algorithm (Forney, 1972), which allows them to compute simultaneously optimal segmentation and classification of the audio stream: noise in individual frames is alleviated by the HMM's learned hysteresis, i.e., its typical learned preference for self-transitions rather than non-self-transitions in the hidden finite state machine.

To take advantage of this proven approach, we leverage a framework in which HMMs are used to achieve audio segmentation and event classification simultaneously. To alleviate HMM's problem that each hidden state models only local observations, we propose to use the tandem connectionist-HMM approach (Hermansky et al., 2000), where an artificial neural network (ANN) outputs posterior probabilities of event types based on very-long-duration, temporally overlapping observation vectors, leading to better contextual modeling and event discrimination. To further refine the event detection result, we propose using vectors of the per-segment adapted means of a Gaussian mixture model (GMM), so-called GMM supervectors (Campbell et al., 2006), to abstract the noisy features in the training audio segments and the hypothesized segments obtained by the tandem model. An SVM with kernels built on these GMM supervectors, namely the SVM–GMM-supervector classifier, is used to replace the labels proposed by the first-pass tandem model, when such replacement is desirable according to held-out development data.

We perform acoustic event detection experiments on the same setup as the AED Evaluation in CLEAR 2007. It is demonstrated that the discriminative feature set constructed by the boosted feature selection approach, the tandem connectionist-HMM approach and the SVM–GMM-supervector approach for refining the result jointly contribute to performance improvement from 28.2% to 41.2% absolute. This also outperforms our submission in the CLEAR 2007 AED Evaluation, which was the best ranked in the challenging AED task.

2. Discriminative features for AED

2.1. Spectral correlates of acoustic events

Over the past decades, a lot of research has been done on speech perceptual features (Hermansky, 1999; Reynolds and Rose, 1995). Currently, the speech features are designed mainly based on properties of speech production and perception. Based on knowledge of the human auditory system, the envelope of the spectrogram (formant structure) instead of the fine structure of the spectrogram (harmonic structure) is believed to hold most information for speech. Both log frequency filter bank parameters and Mel Frequency Cepstral Coefficients (MFCC) (Hermansky, 1999) use triangular band pass filters to smooth out the fine structure of the spectrogram. Moreover, to simulate the non-uniform frequency resolution observed in human auditory perception, these speech feature sets use bandwidths based on the perceptual critical band, e.g., they have higher resolution in the low frequency part of the spectrum. These features have been successfully used to characterize speech signal as well as other signal perceived by human audition, e.g., music (Logan, 2000).

The spectral structure of acoustic events is different from that of speech, as shown in Fig. 1, therefore speech feature sets designed according to the spectral structure of speech might be far from optimal for AED. For example, they might neglect frequency ranges that contain little speech discriminative information, but which may contain much discriminative information for acoustic events.

To analyze the spectral structure of acoustic events for AED, we carry out Kullback–Leibler Divergence (KLD) based feature discriminative capability analysis. This helps us to understand the relevance of different feature components (in a speech feature set) for the AED task, compared to speech recognition. The distance between the distributions associated with an acoustic event label and the other audio labels reveals the discriminative capability of the feature for that acoustic event.

KL Divergence (KLD), denoted by $D(p||q)$, is a measure (a "distance" in a heuristic sense) between two distributions, p and q, and is defined as the cross entropy between p and q minus the self entropy of p.

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)}. \quad (1)$$

We use KLD to measure the discriminative capability of each feature component for each acoustic event. Let $d_{ij} = D(p_{ij}||q_i)$ denote the divergence between the distribution of the ith feature component for the jth acoustic event and the global distribution of the ith feature component for all the audio.

The global discriminative capability of the ith feature component is defined by

$$d_i = \sum_j P_j d_{ij}, \quad (2)$$

where P_j is the prior probability for the jth acoustic event.

To calculate the KLD without prior knowledge of each feature component's distribution, we use nonparametric density estimation, in particular, Parzen window density estimation (Duda et al., 2001) with Gaussian kernels to estimate the distribution of each feature component for each event.

The global discriminative capabilities for different log frequency filter bank parameters are estimated for AED and digit classification. The AED data used is the training data used in the detection experiments, as detailed in Section 7. The task of speech digit classification uses digit speech data in TIDIGITS dataset (Tidigits, 1993). In these preliminary experiments, we observe that the tasks of spoken digit recognition and acoustic event detection assign different relative levels of importance to each of the feature components.

2.2. Boosted feature selection

As discussed in the above subsection, the sum of the KLD between every event-specific distribution and the global distribution characterizes the discriminative capability of the concerned feature component. The goal of feature selection, however, is to find the most discriminative feature set instead of finding a set of individually most discriminative feature components.

A few algorithms exist for feature selection. In particular, a floating search approach was proposed in (Pudil et al., 1994), and an extended and more complicated version was later reported in

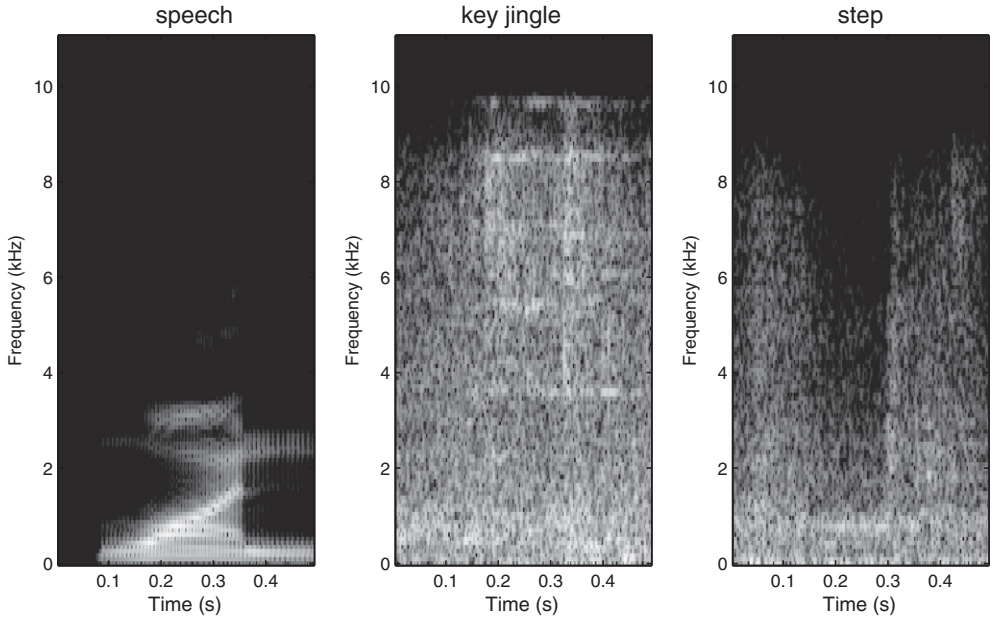


Fig. 1. Spectrograms of the acoustic events "Key Jingle", "Step" and human speech.

(Somol et al., 1999). Jain and Zongker (1997) gave a brief survey of some popular feature selection algorithms. In this work, we use a boosted feature selection approach, inspired by the AdaBoost algorithm (Freund and Schapire, 1999).

The basic AdaBoost algorithm (Freund and Schapire, 1999) is widely used to deal with 2-class classification problems. It iteratively selects and linearly combines several effective classifiers among a lot of weak classifiers.

In this work, the weak classifier selection mechanism of AdaBoost is used to select the feature components. Each audio session in the development set is segmented to acoustic event instances as well as background, according to human transcribed labels. These labeled frames serve as the labeled examples. For each frame, we calculate the ratio of the frame likelihood given the correct acoustic event label vs. the global distribution, where the distributions are estimated using Parzen windows. A boosting approach is then applied to select features: each feature is considered to be a classifier, which labels a frame correctly if its likelihood ratio is greater than one. Note that as in AdaBoost, the chosen "learning rate" α minimizes the normalization term Z_t, which is equivalent to minimizing the training error.

The steps of the boosted feature selection approach are as follows:

1. Prepare the labeled frames $x_1, x_2, \ldots, x_m$, and the corresponding labels $y_1, y_2, \ldots, y_m$.
2. Initialize weights $D_1(i) = \frac{1}{m}, i \in \{1, \ldots, m\}$ where m is the total number of labeled examples.
3. For $t = 1, \ldots, T$, where T is the total number of features:
   (a) Find the feature F_t that minimizes the error ϵ_t with respect to the weights D_t. The error for this iteration is defined as

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)[LLR_{F_t}(x_i) \leqslant 1],$$

   where $[\cdot]$ is the unit indicator function, and $LLR_{F_t}(x_i)$ is the ratio of the likelihoods of feature F_t in frame x_i given the correct event label y_i, vs. the global distribution.
   (b) Choose $\alpha_t \in \mathbf{R}$, set $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$
   (c) Update weights D_t:

$$D_{t+1}(i) = D_t(i) \frac{\exp\{-\alpha_t \cdot sign(LLR_{F_t}(x_i) - 1)\}}{Z_t},$$

   where Z_t is a normalization constant, such that

$$\sum_{i=1}^{m} D_{t+1}(i) = 1.$$

4. Output the first N selected features F_t, where $t \in \{1, \ldots, N\}$.

3. HMM-based AED system

Audio event detection requires both segmentation of the audio stream, and classification of the segments. Following our experience in the AED task of CLEAR 2007, we perform simultaneous segmentation and classification using a Bayesian inference procedure similar to state-of-the-art methods for continuous speech recognition (Ratsch et al., 2001; Freund and Schapire, 1999).

We formulate the goal of acoustic event detection as follows: to find the event sequence that maximizes the posterior probability of the event sequence $W = (w_1, w_2, \ldots, w_M)$, given the observations $O = (o_1, o_2, \ldots, o_T)$:

$$\widehat{W} = \arg\max_W P(W|O) = \arg\max_W P(O|W)P(W). \quad (3)$$

The acoustic model $P(O|W)$ is one HMM for each acoustic event, with three emitting states connected using left-to-right and self-loop transitions. For background silence and speech, we use a HMM with additional transitions between the first and third emitting states, to account for the increased internal complexity. The structure of the HMMs can model some of the non-stationarity of acoustic events. The observation distributions of the states are incrementally-trained Gaussian mixtures. The HMM for an acoustic event is trained to represent all training data segments carrying the same event label.

In order to capture short-term soft constraints on the sequence of event labels, the probability of an event label sequence $(w_1, \ldots, w_m)$ is represented by a bigram language model:

$$P(w_1 w_2, \ldots, w_m) = P(w_1) \prod_{i=2}^{m} P(w_i|w_{i-1}). \quad (4)$$

A bigram "language model" in AED favors recognized acoustic event sequences with sequence statistics similar to those in the training data. Although the language model here does not have the same linguistic implications as in speech recognition, it does improve performance. One of the possible reasons is that it suppresses long sequences of identical event labels in decoding. This is desirable as it forces the HMMs to better fit the internal temporal structure of the audio segments corresponding to the acoustic events.

4. Tandem connectionist-HMM approach

The tandem connectionist-HMM approach is composed of two major components, as shown in Fig. 2: an artificial neural network (ANN) that observes feature vectors in a context window and outputs posteriors of different acoustic event types, and an HMM component that uses a transformed and normalized version of the output of the ANN, optionally together with the original features, as input features. This approach has been shown to improve HMM-based automatic speech recognition (Hermansky et al., 2000). We use the same framework to boost performance of acoustic event detection by drawing evidence from a wider time context window and emphasizing difference between confusable feature vectors across acoustic events by discriminative training.

Two lessons from its application in speech recognition is particularly relevent for using the approach in AED. First, the ANN improves recognition performance in high noise conditions (Ellis and Gomez, 2001; Ellis et al., 2001). The AED task also characterizes low SNR, in particular with background that has high variation. Second, the ANN benefits speech recognition when context independent models are used (Ellis et al., 2001). To limit the complexity of the ANN, it is used to distinguish only between different context-independent models. As pointed out by Ellis et al. (2001), if the generative (HMM) part of the tandem system leverages context-dependent models, the ANN may end up counterproductive by increasing overlap and confusion between different context-dependent models that correspond to the same context-independent model. In this work, we use the HMMs to model different acoustic events that are indeed context-independent.

Consecutive frames within the context window are concatenated to form the input X to the ANN, each dimension corresponding to one input node. The number of output nodes equals the number of acoustic event types. The ANN is discriminatively trained, by back-propagating a minimum cross entropy criterion, to targets that set the output node corresponding to the ground truth event as one and all other output nodes as zero. During testing, for each context window, the ANN presents estimated posterior probabilities across all acoustic events. All context windows centered at every consecutive feature frame are evaluated in the same way, resulting in a sequence of posterior probability vectors.

With these posterior probabilities, we could perform classification using two different approaches. The first approach just directly uses the ANN output: either to assign to each frame its Maximum *A Posteriori* event label, or to generate probabilities that will be smoothed by a Viterbi decoder. However, experiments in automatic speech recognition suggests that better results may be obtained by transforming the posteriors into a pseudo-observation, which is then used as the input to a Gaussian mixture HMM.

In order for ANN posterior probability vectors to be better modeled by the Gaussian mixture likelihood model of an HMM, three transformation are applied as suggested by previous work in tandem speech recognition (Hermansky et al., 2000). First, we take the log of each posterior probability to reduce the skewedness of the distributions. Second, principal component analysis (PCA) is applied on the log probabilities to decorrelate the HMM input, so that we may use diagonal covariance matrices in the Gaussian mixture models. Third, mean and variance normalization is applied on each of the decorrelated dimensions, within each audio session.

5. SVM–GMM-Supervectors

Researchers in automatic speaker identification have recently developed a set of algorithms that boost classification performance by feeding the parameters of a generative model (usually by adapting, to each class, the Gaussian component mean parameters of a universal background GMM) as the input of a discriminative classifier (usually an SVM) (Campbell et al., 2006). The SVM–GMM-supervector approach is not practical as a first-pass segmenter for AED, because it requires some type of hypothesized segment boundaries. Given the boundaries chosen by an connectionist-HMM first-pass system, the SVM–GMM is able to efficiently compute confidence scores for each of the proposed segment labels. The SVM–GMM finely differentiates different candidate classes

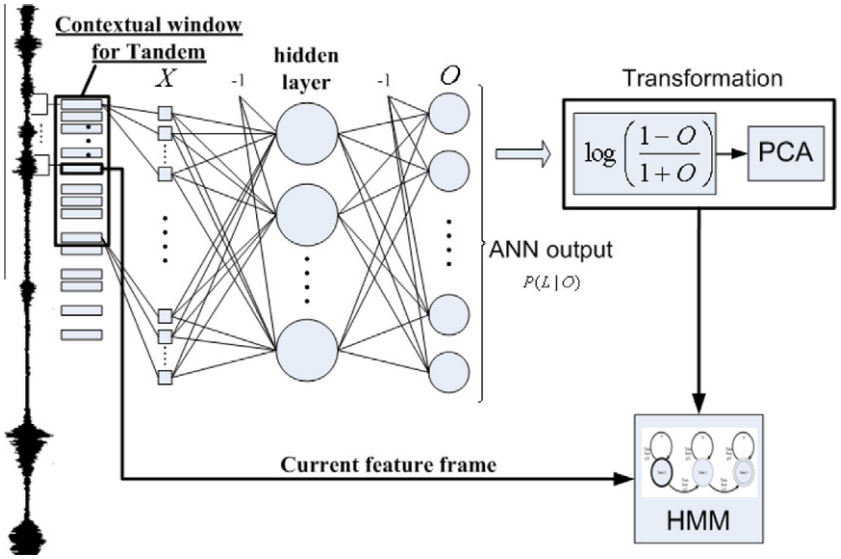


Fig. 2. Classification using a tandem model (ANN + HMM).

by normalizing each class by adaptation of a common multi-mode Gaussian mixture distribution in a discriminative framework.

We refer to the audio observation between two adjacent boundaries as an *audio segment*. The SVM–GMM-supervector approach approximates the joint distribution of all feature vectors in *each audio segment* with a GMM, from which a GMM supervector is constructed as a summary of the segment. The pairwise Euclidean distances between these supervectors characterize the difference between the audio segments. Kernels derived from these distances are used in an SVM for classification. Some of the presentations below follow Campbell et al. (2006), where readers might find more relevant details.

5.1. Universal background model and segment-specific GMM

We estimate a GMM for the distribution of all feature vectors in each audio segment. Instead of separately estimating a GMM for each audio segment, we estimate a GMM for each audio segment by adapting, to each audio segment, the parameters of a universal background model (UBM): a GMM that has been previously trained to represent all types of audio. Adaptive training creates a regularized estimate of the true, underlying likelihood function governing each audio segment. Regularization (adaptative training based on a UBM) reduces the effects of outliers, e.g., noisy frames in an audio segment. Adaptive training also provides a natural measure of the difference between any given audio segment and the UBM, since each Gaussian kernel in the segment-specific likelihood has been adapted from a particular kernel of the UBM. Conversely, the use of a GMM allows arbitrarily precise representation of the acoustic feature likelihood, with large enough number of Gaussian components. Finally, the GMM clusters similar frames, by assigning them to the same kernel in the GMM.

We first estimate a UBM using feature vectors extracted from all training audio segments, regardless of their event labels. Then the distribution model of the feature vector for a certain audio segment is adapted from the UBM in order to maximize the *a posteriori* probability of the adapted model (Gauvain and Lee, 1994).

Here we denote $z \in \mathbb{R}^d$ as a feature vector, where d is the dimension of the feature vector. The GMM distribution of variable z is

$$p(z;\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(z;\mu_k,\Sigma_k), \tag{5}$$

where $\Theta = \{w_1,\mu_1,\Sigma_1,\ldots\}, w_k, \mu_k$ and Σ_k are the weight, mean, and covariance matrix of the kth Gaussian kernel, respectively, and K (set as 128 in this work) is the total number of Gaussian kernels.

The density is a weighted linear combination of K unimodal Gaussian densities, namely,

$$\mathcal{N}(z;\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}. \tag{6}$$

We obtain maximum likelihood parameters for the UBM using Expectation–Maximization (EM). For computational efficiency, the covariance matrices are restricted to be diagonal, which proves to be effective and computationally economical.

The UBM, learnt from all training audio, specifies a feature domain, of which each segment-specific GMM span a subset. The subset constraint can be enforced by interpreting the UBM parameter set, Θ, as a set of conjugate-prior PDFs governing the distribution of segment-specific GMM parameters, θ, i.e., the segment-specific GMM has the *a priori* PDF $p(\theta;\Theta)$ (Lee et al., 1991). The *a posteriori* probability of the segment-specific GMM parameters is obtained by multiplying $p(\theta;\Theta)$ by the data likelihood, $p(Z|\theta)$, where $Z = \{z_1,\ldots,z_H\}$ are the frames observed belonging to the segment of interest, and by then dividing by a normalizing constant; the normalizing constant is irrelevant to computation of the model parameters, and may be omitted. Thus, for example, MAP adaptation selects the segment-specific mean parameters $\hat{\mu}_k$ to maximize

$$\ln p(\hat{\theta},Z) = \sum_{k=1}^{K} \ln \mathcal{N}(\hat{\mu}_k;\mu_k,\Sigma_k/r) + \sum_{i=1}^{H} \ln \sum_{k=1}^{K} w_k \mathcal{N}(z_i;\hat{\mu}_k,\Sigma_k), \tag{7}$$

where $\hat{\theta} = \{\hat{\mu}_1,\ldots,\hat{\mu}_K\}$ is the set of segment-specific GMM parameters, $\Theta = \{w_1,\mu_1,\Sigma_1,\ldots\}$ are the parameters of the global GMM, and r is a regularization constant.

The joint distribution function $p(\hat{\theta},Z)$ has the same form as the likelihood function $p(Z|\hat{\theta})$, and may therefore be optimized in the same way as a likelihood function, *i.e.*, using EM with the hidden variable $Pr(k|z_i)$ as the posterior probability of the Gaussian component k for given feature vector z_i (Lee et al., 1991). In the E-step, we compute the posterior probability as

$$Pr(k|z_i) = \frac{w_k \mathcal{N}(z_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(z_i;\mu_j,\Sigma_j)}, \tag{8}$$

$$n_k = \sum_{i=1}^{H} Pr(k|z_i) \tag{9}$$

and then the M-step updates the mean vectors, namely,

$$E_k(Z) = \frac{1}{n_k}\sum_{i=1}^{H} Pr(k|z_i) z_i, \tag{10}$$

$$\hat{\mu}_k = \alpha_k E_k(z) + (1-\alpha_k)\mu_k, \tag{11}$$

where $\alpha_k = n_k/(n_k+r)$. MAP adaptation using conjugate priors is useful because it interpolates, smoothly, between the hyper-parameters μ_k and the maximum likelihood parameters $E_k(Z)$. In this work, r is adjusted, empirically, depending on the total number of feature vectors for each audio segment.

5.2. Approximating Kullback–Leibler divergence

Two *segment-specific* GMMs adapted from the same UBM are denoted as g_a and g_b. A natural similarity measure between these two GMMs is the Kullback–Leibler divergence,

$$D(g_a||g_b) = \int_z g_a(z)\log\frac{g_a(z)}{g_b(z)}\, dz.$$

The Kullback–Leibler divergence does not satisfy the conditions for a metric function. Instead, we can use its upper bound obtained by the log-sum inequality,

$$D(g_a||g_b) \leqslant \sum_{k=1}^{K} w_k D\big(\mathcal{N}(z;\mu_k^a,\Sigma_k)||\mathcal{N}(z;\mu_k^b,\Sigma_k)\big),$$

where μ_k^a and μ_k^b denote the adapted means of the kth component from the segment GMMs g_a and g_b, respectively. Since the covariance matrices are shared across all adapted GMMs and the UBM, the right hand side is equal to

$$d(a,b)^2 = \frac{1}{2}\sum_{k=1}^{K} w_k\big(\mu_k^a-\mu_k^b\big)^T\Sigma_k^{-1}\big(\mu_k^a-\mu_k^b\big).$$

We can consider $d(a,b)$ as the Euclidean distance between the normalized GMM supervectors in a high-dimensional feature space (Zhou et al., 2008),

$$d(a,b) = \|\phi(Z_a)-\phi(Z_b)\|_2, \tag{12}$$

where

$$\phi(a) = \left[\sqrt{\frac{w_1}{2}}\Sigma_1^{-\frac{1}{2}}\mu_1^a;\cdots;\sqrt{\frac{w_K}{2}}\Sigma_K^{-\frac{1}{2}}\mu_K^a\right]. \tag{13}$$

5.3. Kernel for SVM

GMM supervectors are used in an SVM for acoustic event classification. This multi-class classification task is implemented as binary classification problems via the one-vs-one method using LibSVM (Chang and Lin, 2001). The distance defined in (12) can be evaluated using kernel functions, as

$$d(a,b) = \sqrt{K(a,a) - 2K(a,b) + K(b,b)}. \quad (14)$$

It is straightforward that kernel function $K(a,b) = \phi(a) \cdot \phi(b)$ satisfies (14), where $\phi(a)$ and $\phi(b)$ are defined as in (13).

6. Hybrid architecture of AED system

Both the HMM-based approach and the tandem HMM-connectionist approach engage the maximum *a posteriori* probability (MAP) decoding for AED, the recognizer outputs a sequence of hypothesized acoustic events corresponding to the highest *sequence* a posterior probability, as discussed in Section 3. However, the best acoustic event sequence obtained by the MAP decoding is not optimal according to the performance measure for AED, *AED-ACC*, i.e. the acoustic event F-score (harmonic mean of precision and recall). For example, Mangu et al. (2000) proposed solving a similar problem using localized confidence rescoring: the MAP decoder defines a reduced search space, within which a new hypothesis is chosen explicitly to minimize the target performance measure. Confidence scoring also allows us to apply methods such as SVM–GMM-supervector classification, which are difficult to apply in a MAP decoding paradigm because of computational complexity and model structure limitations.

In this work, our final system uses a two stage hybrid architecture (Fig. 3). In (Mangu et al., 2000) a rescoring paradigm aligns all of the edges in an event lattice to the times marked in the MAP hypothesis. In the AED task, the number of labels is small enough to obviate lattice rescoring, therefore we can take a route that is straightforward, yet effective and computationally inexpensive. The MAP decoding outputs a one-best result with boundaries of events and background, as well as hypothesized event types. The SVM–GMM-supervector approach is used as the confidence rescoring module. It models feature frames within all hypothesized audio segments, and proposes event types that might be different from the hypothesis obtained through MAP decoding.

Both hypothesized event types, referred to as the MAP labels and the SVM labels, respectively, include the events of concern and a "background" label. Therefore, event label substitutions, each defined by a MAP label and an SVM label, may include substitutions between any pair of events, from an acoustic event to background or from background to an acoustic event. On the held out development data, the performance change is measured when only one particular type of label substitution is allowed or not. Those label substitution types that lead to the most performance boost on the held out data are chosen as the *valid event label substitutions*, to be applied in testing.

We find in practice that the above valid event label substitutions are too specific and sometimes do not carry over well between different data. Therefore, in the experiments we only define valid event label substitutions according to the MAP labels. In fact, the most favorable approach turns out to allow the SVM–GMM-supervector classifier to assign labels to the audio segments labeled as background by the MAP decoding, recovering events that were missed in the first pass.

We speculate that the hybrid architecture might work for two reasons.

First, the SVM–GMM-supervector approach functions complementary to the MAP decoding as they operate in different hypothesis spaces. In particular, the MAP decoding engages properties such as state transition, varying length and *N*-gram event sequence statistics in the decision of boundaries and hypothesized event labels. The MAP decoding might suppress proposing short events or events similar to the background given the high variation in the background. By contrast, the SVM–GMM-supervector approach only considers feature distribution within an audio segment locally. The purely local approach of the rescoring module has been shown to outperform HMMs in tasks with loose sequence constraints (Huang et al., 2009).

Second, the objective of MAP decoding differs from that of AED. For maximum *a posteriori* hypothesis, each frame in the observation is considered. The detection metric *AED-ACC*, only considers the relative time relationship among the hypothesized event boundaries. Furthermore, neither MAP decoding nor the SVM–GMM-supervector classifier treat background and acoustic events differently, while the *AED-ACC* measures only the *F*-score in detection of non-background events. SVM–GMM rescoring aims at the target performance metric by constraining it to allow only label substitutions (changes from the MAP labels) that are believed to improve the AED performance metric.

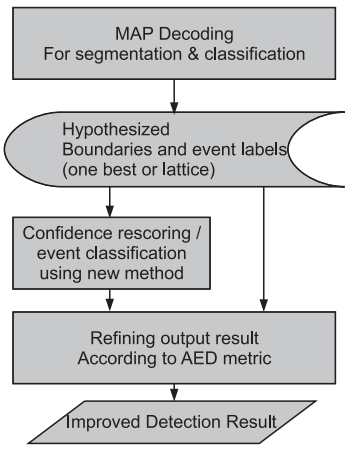


Fig. 3. Hybrid architecture of AED system.

7. Experiments

7.1. Dataset and metric

Acoustic event detection experiments use the official data for CLEAR 2007 AED Evaluation (Temko, 2007): about three hours for system development and two hours for system evaluation. All data are realistic seminar style, having both speech and acoustic events with possible overlap. The evaluation data has 1454 instances of target events. The target events included in the AED performance metric are door slam (ds), paper wrapping/rustling (pw), footsteps (st), phone ringing (pr), spoon/cup jingle (cl), keyboard typing (kt), applause (ap), coughing (co), laughter (la), key jingle (kj), chair moving (cm), and knocking (kn). The histogram of these events in the evaluation data is as in Fig. 4. Many of the events are subtle and have low SNR compared to background noise or speech. The non-target labels in the dataset include an unknown event label and a speech label. In this work, both unlabeled frames and frames labeled as speech are treated as the background class.

As mentioned in Section 6, the performances are measured using *AED-ACC* (Temko, 2007), defined as the *F*-score (the harmonic mean between precision and recall) comparing system output acoustic event (AE) labels and reference AE labels. In

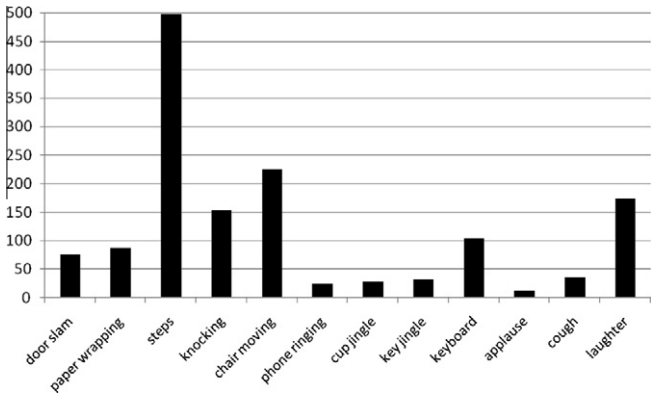


Fig. 4. Histogram of the twelve events in the evaluation data.

particular, an event detected by the system is correct when there exist at least one matching reference event whose temporal center falls within the time boundaries of the detected event or the temporal center of the detected event is within the boundaries of at least one matching reference event. A reference event is considered correctly detected if its temporal center is within at least one matching system output or if there exist at least one matching system output whose temporal center falls within the boundaries of the reference event. *AED-ACC* aims to score detection and classification of all acoustic event instances, oriented for applications such as real-time services for smart rooms and audio-based surveillance.

7.2. Experiment set-up

Three sets of experiments are carried out to demonstrate the performance of the derived features, the tandem connectionist-HMM approach and the SVM–GMM-supervector approach for refining event label hypotheses.

The first experiment compares the performance of the HMM-based AED systems using either the derived AED feature or the baseline set MFCC or log frequency filter bank parameters. Both baseline feature sets are widely-used in speech recognition as well as other audio applications. The first baseline feature set consists of 26 MFCCs calculated in the 0–11,000 Hz band along with their first order regression (delta) coefficients and second order regression (acceleration) coefficients. The second baseline feature set consists of 26 log frequency filter bank parameters, their delta and acceleration coefficients on the same frequency range. The AED feature set is derived using the boosting approach discussed in Section 2, from the union of the two baseline feature sets. Each feature set used in this experiment has 78 feature components, and each system has the same number of trainable parameters.

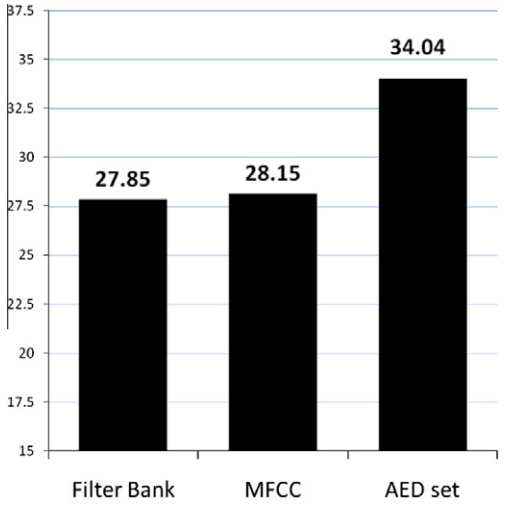
The second experiment evaluates the tandem connectionist-HMM approach. The contextual window size (number of input nodes divided by 78) is picked to be five. The number of hidden nodes is chosen as 1200 empirically for best performance on a development dataset. The number of output nodes is set to 14, i.e., the number of acoustic events plus one for frames labeled as unknown sounds and one for background frames. The transformed output of the best-performing ANN is concatenated with the derived AED feature set as the input to the HMM component.

The third experiment presents performance of the SVM–GMM-supervector approach discussed in Section 5, used in the hybrid architecture discussed in Section 6. The number of Gaussian mixtures is set to be 128. Two sets of results are reported, obtained by applying the approach on top of either the HMM-based approach or the tandem connectionist-HMM approach.

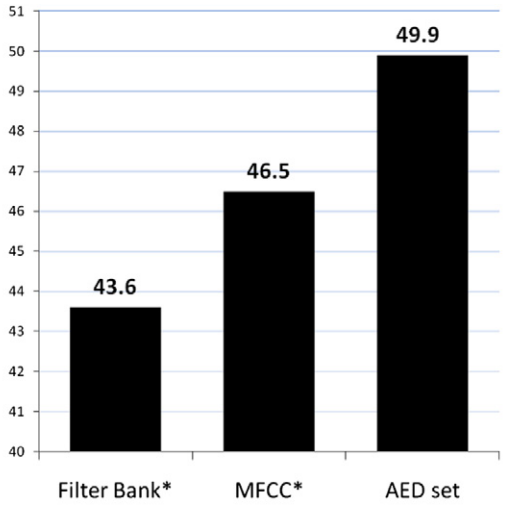
When training the systems, we hold out one third of the three hours development data to tune some system parameters. Once the parameters are determined, the models are retrained with all the development data.

7.3. Experiment results

Fig. 5 presents the performance of the AED feature set derived using the boosting approach, and the baseline MFCC or log frequency filter bank parameters. The detection accuracy, measured by *AED-ACC* score, is presented in Fig. 5a. It is shown that the



(a) *AED-ACC* scores %



(b) Classification accuracy % (* indicates significantly different from AED set (McNemar's test, $p = 0.05$))

Fig. 5. Performance using different features sets.

Table 1
Effectiveness of each components in our framework (*AED-ACC* (%)). Note: (1) 'T' denotes the Tandem approach; (2) 'S' denotes the SVM–GMM-Supervector approach.

Frequency	ap	cl	cm	co	ds	kj	kn	kt	la	pr	pw	st	Average
	13	28	226	36	76	32	153	105	174	25	88	498	121
MFCC	**78.3**	**26.9**	29.5	24.2	56.3	**39.9**	7.7	0.0	39.0	35.2	14.1	28.7	28.2
FB	34.5	21.8	25.4	24.9	38.9	27.2	11.7	0.0	49.1	13.8	11.7	28.1	27.8
Boosted	44.4	25.5	31.3	31.2	57.3	33.2	13.5	1.9	51.3	36.7	17.6	36.8	34.0
Boosted + T	52.6	21.9	37.2	**51.3**	**63.0**	29.6	11.5	0.0	54.2	**42.7**	25.8	34.6	35.3
Boosted + S	44.4	25.0	33.7	31.2	56.6	33.2	**20.9**	35.5	51.3	36.7	19.2	41.3	37.5
Boosted + T + S	52.6	21.5	**37.4**	47.9	**63.0**	29.6	13.6	**44.8**	**58.6**	**42.7**	**26.7**	**44.4**	**41.2**

AED feature set outperforms both MFCC and log frequency filter bank features without dimension increase.

It is not straightforward to carry out a statistical test on the detection metric. To verify that the derived AED feature set does perform significantly different from the baseline speech feature sets, we present a classification experiment (Fig. 5b). The HMMs in the AED systems are used to classify audio segments into different events as well as background. These audio segments include those acoustic events extracted according to the ground truth of the testing data, as well as background segments, which are areas that do not have any event label. The classification results using different feature sets are subject to a McNemar's test. The performances of both MFCC and log frequency filter bank parameters differ significantly from the derived AED feature set at the 95% confidence level.

In Table 1, we demonstrate the effectiveness of the tandem HMM-connectionist approach and the SVM–GMM-supervector approach used in the hybrid architecture. We can observe that the average *AED-ACC* across all twelve events improves from 34% to 35.3% by engaging the tandem approach (denoted as "Boosted + T"). The SVM–GMM-supervector (denoted as "Boosted + S") boosts performance from 34% to 37.5% by relabeling event segments proposed by the HMM-based AED system, as described in Section 6. Using this hybrid architecture of both tandem and SVM–GMM-supervector approaches yields the best *AED-ACC* of 41.2% (denoted as "Boosted + T+S"). The frequencies of the events are also included for reference.

The best performance with the AED feature set, the tandem HMM-connectionist approach and the SVM–GMM-supervector approach is over 45% relative improvement from the baseline feature and model (MFCC + HMM 28.2% or filter bank + HMM 27.8%), and outperforms the best previously reported performance on CLEAR AED task.

Performance on individual acoustic events is also presented for the different settings. It is shown that the number of individual acoustic events scoring the highest is the largest for the best setting of "Boosted + T+S". The single most dramatic performance boost on an individual event is that of "keyboard typing" (kt), achieved by engaging the SVM–GMM-supervector approach. The MAP decoding approaches, i.e., HMM or tandem approaches, could not well distinguish "keyboard typing" from background. In fact, many events that are easily confused with the background in the first pass, e.g., "keyboard typing" and "steps", are recovered for reasons discussed in Section 6. This highlights that the SVM–GMM-supervector in the hybrid architecture has capability complementary to the MAP decoding approaches.

8. Conclusion

In this paper, we present both discriminative features and system architectures designed for better acoustic event detection. We propose a boosting approach to derive the most discriminative feature set for AED from a feature pool of MFCC and log frequency filter bank parameters. Inspired by advances in speech recognition, a tandem connectionist-HMM approach for AED is proposed to combine the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained using the minimum cross entropy criterion. Finally, an SVM–GMM-supervector approach is designed using noise-adaptive kernels better approximating the KL divergence between feature distributions in different audio segments. Experiments on the CLEAR AED Evaluation set-up demonstrate that the presented features and models all contribute toward improved performance, compared with previous best-performing approaches (Zhou et al., 2007), on detection of twelve general acoustic events in a real seminar environment.

Acknowledgement

This work was funded by NSF Grants 08-03219 and 08-07329. The results and conclusions expressed in this paper are those of the authors, and are not endorsed by the NSF.

References

Atrey, P.K., Maddage., N.C., Kankanhalli, M.S., 2006. Audio based event detection for multimedia surveillance. In: ICASSP06.
Baillie, M., Jose, J., 2003. Audio-based event detection for sports video. Lecture Notes Comput. Sci. 2728, 61–65.
Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8, 297–336.
Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A., 2006. SVM based speaker verification using a GMM supervector kernel and nap variability compensation. ICASSP 2006, vol. 1. IEEE, pp. 97–100.
Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
Clavel, C., Ehrette, T., Richard, G., 2005. Events detection for an audio-based surveillance system. In: IEEE Internat. Conf. on Multimedia and Expo., pp. 1306–1309.
Cui, R., Lu, L., Zhung, H.-J., Cai, L.-H., 2003a. Highlight sound effects detection in audio stream. In: ICME03, pp. III: 37–40.
Cui, R., Lu, L., Zhung, H.-J., Cai, L.-H., 2003b. Highlight sound effects detection in audio stream. In: ICME03, pp. III: 37–40.
Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley & Sons, New York.
Ellis, D., 1996. Prediction–driven computational auditory scene analysis. Ph.D. Thesis, MIT.
Ellis, D., Gomez, M.R., 2001. Investigations into tandem acoustic modeling for the aurora task. In: Proc. Eurospeech-01. ISCA, pp. 189–192.
Ellis, D., Singh, R., Sivadas, S., 2001. Tandem acoustic modeling in large-vocabulary recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, 2001 (ICASSP '01), vol. 1. pp. 517–520.
Forney, G.D., 1972. Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. IEEE Trans. Inform. Theory 18 (3), 363–378.
Freund, Y., Schapire, R.E., 1999. A short introduction to boosting. J. Japanese Soc. Artif. Intell. 14 (5), 771–780.
Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markovchains. IEEE Trans. Speech Audio Process. 2, 291–298.
Hermansky, H., 1999. Mel cepstrum, deltas, double deltas, – what else is new? In: Proc. Robust Methods for Speech Recognition in Adverse Condition.
Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature stream extraction for conventional HMM systems. ICASSP 2000, vol. III. IEEE, pp. 1635–1638.
Huang, J., Zhuang, X., Libal, V., Potamianos, G., 2009. Long-time span acoustic activity analysis from far-field sensors in smart homes. In: ICASSP 2009. IEEE.
Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Machine Intell. 19, 153–158.

Lee, C.-H., Lin, C.-H., Juang, B.-H., 1991. A study on speaker adaptation of the parameters of continuous density hidden markov models. IEEE Trans. Signal Process. 39 (4), 806–814.

Logan, B., 2000. Mel frequency Cepstral coefficients for music modeling. In: Proc. Internat. Conf. on Music Information Retrieval.

Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. Comput. Speech Lang. 14 (4), 373–400.

Naphade, M.R., Garg, A., Huang, T., 2001. Duration dependent input output markov models for audio-visual event detection. In: ICME01, p. 65.

Pinquier, J., 2002. Robust speech/ music classification in audio document. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), pp. III: 2005–2008.

Pudil, P., Ferri, F., Novovicova, J., Kittler, J., 1994. Floating search methods for feature selection with nonmonotonic criterion functions. In: Proc. 12th IAPR Internat. Conf. on Pattern Recognition, 1994. Conference B: Computer Vision and Image Processing, vol. 2. pp. 279–283.

Ratsch, G., Onoda, T., Muller, K.-R., 2001. Soft margins for AdaBoost. IEEE Trans. Signal Process. 42, 287–320.

Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3 (1), 72–83.

Scheirer, E.D., 1999. Sound scene segmentation by dynamic detection of correlogram comodulation. Tech. Rep. 491, MIT Media Laboratory Perceptual Computing Section.

Schòlkopf, B., Smola, A., 2002. Learning with Kernels. MIT Press, Cambridge, MA, US.

Smith, J.A., Earis, J.E., Woodcock, A.A., 2006. Establishing a gold standard for manual cough counting: Video versus digital audio recordings. Cough 2 (6), 1–6.

Somol, P., Pudil, P., Novovičová, J., Paclik, P., 1999. Adaptive floating search methods in feature selection. Pattern Recognition Lett. 20 (11–13), 1157–1163.

Temko, A., 2007. CLEAR 2007 AED evaluation plan and workshop. <http://isl.ira.uka.de/clear07>.

Temko, A., Nadeu, C., 2005. Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing, vol. V. pp. 505–508.

Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M., 2006. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. IV Jornadas en Tecnologia del Habla November.

Tidigits, 1993. Linguistic Data Consortium Catalog No. LDC93S10.

Zhang, T., Kuo, C.-C.J., 2001. Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans. Speech Audio Process. 9 (4), 441–457.

Zhou, X., Zhuang, X., Liu, M., Tang, H., Hasegawa-Johnson, M., Huang, T., 2007. HMM-based acoustic event detection with AdaBoost feature selection. In: Classification of Events, Activities and Relationships Evaluation and Workshop, pp. 345–353.

Zhou, X., Navrátil, J., Pelecanos, J.W., Ramaswamy, G.N., Huang, T.S., 2008. Intersession variability compensation for language detection. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing.

Zhuang, X., Zhou, X., Huang, T.S., Hasegawa-Johnson, M., 2008. Feature analysis and selection for acoustic event detection. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, 2008 (ICASSP '08), pp.17–20.