CLASSIFICATION OF HUMAN COUGH SIGNALS USING SPECTRO-TEMPORAL GABOR FILTERBANK FEATURES

Jens Schröder[1,2]*, Jörn Anemüller*[2,3] *and Stefan Goetze*[1,2]

[1] Fraunhofer IDMT / Hearing, Speech and Audio Technology, 26129 Oldenburg, Germany
[2] Cluster of Excellence Hearing4all, Germany
[3] University of Oldenburg, Dept. of Physics and Acoustics, 26111 Oldenburg, Germany
{jens.schroeder, s.goetze}@idmt.fraunhofer.de, joern.anemueller@uni-oldenburg.de

ABSTRACT

This contribution investigates the use of features derived from a Gabor filterbank (GFB) for the application of acoustic cough classification. Gabor filters are two-dimensional filters that decompose the spectro-temporal power density further into components which capture spectral, temporal and joint spectro-temporal modulation patterns.

The proposed GFB feature extraction scheme in combination with Gaussian mixture model (GMM) and hidden Markov model (HMM) classifier back-ends is evaluated using a cough database recorded by a phone hotline. The database is composed of two kind of coughs, i.e., *dry* and *productive cough*, and other sounds, e.g. speech. Based on these data, we show that GFB features result in better recognition performance than the common Mel-frequency cepstral coefficient (MFCC) baseline for the given task of cough classification. Furthermore, results indicate that GMMs are preferable to HMMs for this kind of data.

Index Terms— cough classification, acoustic event classification, spectro-temporal filters, Gabor filterbank

1. INTRODUCTION

Cough is a reflex to clear the respiratory tracts from mucus and foreign particles. An increased number of coughs is usually an indicator of a respiratory disease, e.g., a cold, influenza etc. During a cold, two types of cough are common: productive and dry cough. An infection in the bronchia results in an increased production of viscous mucus that cannot be removed by the usual ways. Instead, it has to be coughed up resulting in sputum. Therefore, this process is called productive cough. Commonly, periods of productive cough are preceded by periods of dry cough. This kind of cough is commonly caused by a defense mechanism to a virus attacking mucous membranes. Messengers are emitted that stimulate sensitive nerve fibers. Though no mucus or intrusive substances have to be removed from the respiratory system, a coughing reflex is provoked. Thus, dry cough has no functionality and by its violent character can even harm the affected mucous membranes further. Hence, an automatic surveillance system that automatically detects and classifies productive and dry coughs, which is proposed in this contribution, can be beneficial for monitoring the health state of patients in hospitals as well as in in-home care.

Several proposals for acoustic event detection (AED) systems implicitly recognize cough amongst other events [1–4]. Other publications explicitly focus on acoustic cough classification. In [5], hidden Markov models (HMMs) and Mel-frequency cepstral coefficients (MFCCs) are proposed for cough detection. In [6], several low-level features and different feature-selection algorithms were tested for cough detection. Even for non-human coughs, automatic detectors were examined [7].

The acoustic characteristics of productive and non-productive coughs was analyzed in [8]. Three temporal phases for coughs were identified. The second phase of productive and dry cough differentiates by its spectral and temporal compositions. Thus, we propose the use of spectro-temporal Gabor filterbank (GFB) features for classification of dry and productive coughs. Gabor filters are two-dimensional spectro-temporal filters. Hence, they are capable of detecting temporal and spectral changes at the feature level jointly in contrast to, e.g., frame-based features like MFCCs. They have been proposed as a mathematical description of spectro-temporal receptive field (STRF) in the auditory processing stages of animals [9–12]. Their use for acoustic feature extraction has been proposed by [13] and [14] in the context of robust automatic speech recognition (ASR). Recently, the features have been applied to AED tasks in [3, 15, 16], resulting in performance increases compared to a MFCC baseline. Other, related methods exploiting two-dimensional (spectro-temporal) context for AED are based on, e.g., spectrogram images [17], stabilized auditory images [18], part-based models [19] or non-negative matrix factorization (NMF) [20, 21]. However, these approaches will not be examined in this con-

This work has been partly funded by the European Commision (project EcoShopping, no. 609180).

tribution.

In the present study, GFB features are used to classify dry and productive coughs and in discrimination against other sounds. We use GFB parameters that have been adopted in [16] for the AED task. For training and testing, real data has been collected by volunteers via a phone hotline. The results will be compared to common MFCC features. Performance difference using Gaussian mixture models (GMMs) and HMMs is investigated.

2. GABOR FILTERBANK

Gabor filters are two-dimensional patterns that are supposed to copy receptive fields of neurons. The GFB employed here is composed of a set of two-dimensional Gabor filters, each defined by its specific temporal and spectral envelope functions and by its temporal and spectral carrier functions, respectively. The Gabor filter as a function of frequency index m and frame index ℓ, with carrier frequency m_0 and temporal frame position ℓ_0, spectral and temporal modulation frequencies ω_m and ω_ℓ and the numbers of semi-cycles under the envelope ν_m and ν_ℓ, respectively, is, thus, defined as

$$\begin{aligned}\gamma\left(m,\ell;m_0,\ell_0,\omega_m,\omega_\ell,\nu_m,\nu_\ell\right)\\ = s_{\omega_m}\left(m-m_0\right)\cdot s_{\omega_\ell}\left(\ell-\ell_0\right)\\ \cdot h_{\frac{\pi\nu_m}{\omega_m}}\left(m-m_0\right)\cdot h_{\frac{\pi\nu_\ell}{\omega_\ell}}\left(\ell-\ell_0\right),\end{aligned} \quad (1)$$

with carrier function

$$s_{\omega_x}(x)=\exp\left(j\omega_x x\right), \quad (2)$$

and envelope function

$$h_b\left(x\right)=\begin{cases}0.5+0.5\cos\left(\frac{2\pi x}{b}\right), & -\frac{b}{2}<x<-\frac{b}{2},\\ 0, & \text{otherwise},\end{cases} \quad (3)$$

where b denotes the filter width. An example of a two-dimensional Gabor filter is depicted in the central pattern of Fig. 2.

The filterbank is designed to cover the spectro-temporal modulation domain approximately uniformly. The spectral and temporal modulation center-frequencies $\omega_m^1,\ldots,\omega_m^{N_m}$ and $\omega_\ell^1,\ldots,\omega_\ell^{N_\ell}$, with N_m and N_ℓ representing the respective number of of center frequencies, are defined recursively according to

$$\omega_m^{i+1}=\omega_m^i\frac{1+\frac{c_m}{2}}{1-\frac{c_m}{2}} \quad \text{with} \quad c_m=d_m\frac{8}{\nu_m} \quad (4)$$

and

$$\omega_\ell^{i+1}=\omega_\ell^i\frac{1+\frac{c_\ell}{2}}{1-\frac{c_\ell}{2}} \quad \text{with} \quad c_\ell=d_\ell\frac{8}{\nu_\ell} \quad (5)$$

with properly chosen lower and upper limits $\omega_m^{\min}$, $\omega_m^{\max}$, $\omega_\ell^{\min}$ and $\omega_\ell^{\max}$. Parameters d_m and d_ℓ define the relative distances of adjacent filters where smaller values correspond to larger filter overlaps.

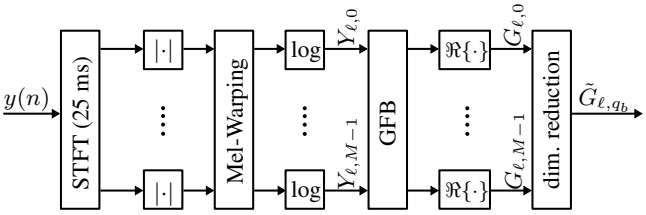

Fig. 1. Block diagram of GFB feature extraction. The input signal is transformed to a logarithmically scaled Mel-spectrogram and decomposed by two-dimensional Gabor filters, yielding high-dimensional GFB features. The dimensionality of the real parts of these features is reduced by a filter function that preserves the most informative feature dimensions while decreasing the number of features for classification.

While purely spectral filters ($\omega_\ell = 0$) are sensitive to spectral patterns like tonal components, purely temporal filters ($\omega_m = 0$) are sensitive to broad-band onsets. Spectro-temporal filters ($\omega_\ell > 0$ and $\omega_m \neq 0$), in contrast, produce highest output when the corresponding joint spectral and temporal transient is observed in the signal. The optimal parameter set for the GFB for the task of acoustic event classification (AEC) has been investigated in [16].

3. FEATURE EXTRACTION

For classification, features are extracted based on the log-scaled Mel-spectrogram $Y_{\ell,k}$ of the signal $y(n)$ with k and n denoting the Mel-band and discrete time index, respectively. The complex-valued Gabor filterbank is applied to the log-scaled Mel-spectrogram $Y_{\ell,m}$ and the output's real part is used for classification, i.e.,

$$\begin{aligned}&G_{\ell,m}\left(m_0,\ell_0,\omega_m,\omega_\ell,\nu_m,\nu_\ell\right)\\ &=\Re\left\{\sum_{\mu,\lambda}Y_{\lambda,\mu}\gamma\left(\mu+m,\lambda+\ell;m_0,\ell_0,\omega_m,\omega_\ell,\nu_m,\nu_\ell\right)\right\},\end{aligned} \quad (6)$$

Applying the GFB to all Mel-bands results in a high-dimensional feature representation, e.g., a Mel-spectrogram with 23 Mel-bands and a GFB with 50 filters result in 1150-dimensional features. Therefore, only Mel-bands that are shifted about $1/4$ of the Gabor filter spectral width and, thus, include substantial new information, are used (cf. Fig. 2). Hereby, feature dimensionality is reduced by a factor of about $1/3$ from 1150 to 380 dimensions. For more details cf. [16]. The procedure of GFB feature extraction is depicted in Fig. 1.

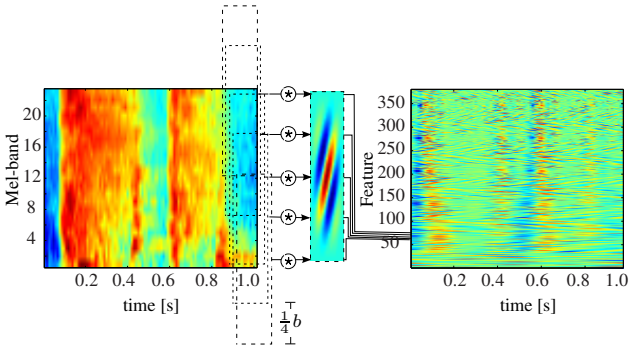

Fig. 2. Illustration of the generation of GFB features (right pattern) from a Mel-spectrogram (left pattern) of a *dry cough* by a single Gabor filter (central pattern) of the filterbank. Dimension reduction is achieved by applying the filter to a subset of central Mel-bands (here: five) that are shifted about $1/4$ of the Gabor filter spectral band width b. (Temporal shifts of the dashed boxes indicating filter positions are just for visualization.)

4. EXPERIMENTAL DATA AND SETUP

Since *productive cough* is hardly producible without being sick, acoustic cough data were collected via a public telephone hotline to gain a sufficient amount of events. Volunteers could call the hotline to check their cough type by coughing into the receiver. Thus, a realistic database with many different participants in various acoustic environments could be gained. The data were recorded at a sampling frequency of $f_s = 8$ kHz and labeled by two human annotators. The annotators labeled each call with one label according to *productive cough*, *dry cough* and other sounds grouped as *garbage* and with leading and tailing silences. The *garbage* class consists of speech (mostly), laughing, music etc. The annotators had the possibility to indicate whether they were sure or unsure with a label. While *garbage* was labeled identically by both annotators, the consent for *dry* and *productive cough* (independent whether sure or not) was only 56.6%. To get reliable ground truth annotations for the following evaluation, only the data that were labeled "sure" and that were labeled consistently between both annotators have been processed. This data set comprises 46 minutes of recordings. Details are given in Table 1.

For classification performance evaluation, the data are divided into five disjoint sets with equal number of events per class. These five sets are used to perform a five-fold cross-validation. *Pause*, i.e. segments of silence, and the *garbage* class are modeled by GMMs, i.e., one emitting HMM state, since they exhibit no temporal structure. In a first experiment, *dry* and *productive cough* classes are modeled by GMMs as well. In a second experiment, left-to-right HMMs [22] with three emitting states are adopted to cover the beginning, the middle and the final phase of dry and productive cough according to the results from [8].

Table 1. Number of events and average duration (mean and standard deviation) per class.

	number events	av. duration [s]
dry cough	228	4.09 ± 2.20
productive cough	124	4.36 ± 2.25
garbage	162	4.83 ± 2.63

Table 2. Parameters for the GFB for the task of AED applied for feature generation (cf. [16]).

d_m	d_ℓ	ν_m	ν_ℓ	$\omega_m^{\min}$	$\omega_\ell^{\min}$	$\omega_m^{\max}$	$\omega_\ell^{\max}$
0.3	0.2	3.5	3.5	0.18	0.22	$\pi/2$	$\pi/2$

State observations are modeled by mixtures of diagonal Gaussians. The optimal number of Gaussian mixture components is estimated by averaging the accuracies of all folds and selecting the mixture number with highest accuracy. Classification is done by Viterbi decoding [22]. The grammar allows for *pause*-event-*pause* states, only. The recognition rate of *pause* will not be considered in the presented results. Results will be given as mean accuracies of all five cross-validation trials for the optimal number of Gaussian mixture components.

Since the sampling frequency is $f_s = 8$ kHz, a Mel-filterbank with 23 filters is used for feature extraction. The window size is 25 ms and the hop size 10 ms. For the GFB features, the optimal parameters for the GFB proposed in [16] for the task of AEC are applied. The parameters are given in Table 2. The dimensionality of the GFB features is 380.

For comparison, standard MFCC features are evaluated as baseline. The MFCCs are based on the same Mel-spectrogram decomposition parameters as the GFB features. A pre-emphasis filter is applied to the time-domain signal that reduces low frequency noise components [22]. The first 12 MFCC coefficients and the zeroth coefficient are used. Additionally, derivatives of first (Δ) and second ($\Delta\Delta$) order are concatenated with the MFCCs to capture temporal dynamics. Thus, they comprise a dimensionality of 39 features.

5. RESULTS

For evaluation of cough classification, two experiments are conducted. In the first experiment, all classes (*pause*, *garbage*, *dry cough*, *productive cough*) are modeled by GMMs. In the second experiment, three-states HMMs are used to model the classes *dry cough* and *productive cough* according to the three cough phases identified by [8]. MFCCs and GFB features are tested with these model types.

The accuracies as mean and standard deviation of the cross-validation trials are depicted in Fig. 3. In the left panel,

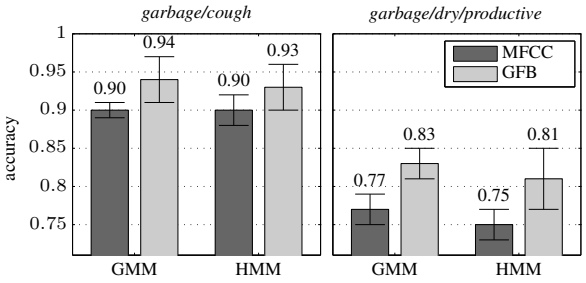

Fig. 3. Accuracies for classification plotted in terms of mean and standard deviation of the five cross-validation trials. In the left panel, accuracies for distinction between *garbage* and *cough* in general are depicted, i.e., confusion of *dry* and *productive cough* are not considered as mistakes. In the right panel, accuracies for classification of *dry cough*, *productive cough* and *garbage* are depicted. Evaluation is based on MFCC (dark grey) and GFB features (light grey) using GMMs for all classes (left side of each panel) and three-states HMMs for models *dry* and *productive cough* (right side of each panel).

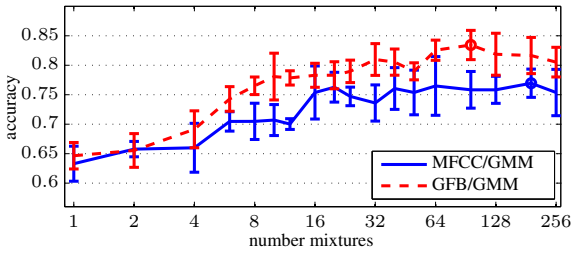

Fig. 4. Mean accuracies and standard deviations (whiskers) from the five-fold cross-validation over the number of mixtures. GMMs are applied based on MFCC (solid) and GFB features (dashed) for classification of *dry cough*, *productive cough* and *garbage*. Maxima are indicated by circles. (Note that the x-scale is logarithmic.)

the mean accuracies and standard deviations of the five-fold cross-validation for distinction of *garbage* and *cough* in general are shown, i.e., classification results of *dry* and *productive cough* are merged and confusions between these two classes are not counted as mistakes. In the right panel, mean accuracies for classification of all three classes are presented, i.e., *garbage*, *dry cough* and *productive cough* are considered as separated, independent classes.

Apparently, GFB features yield higher accuracies for cough classification than MFCCs, irrespectively of the type of modeling. This gain by using GFB features over MFCCs is not an effect of random accuracy maxima originating from model parameterization during cross-validation as can be seen in Fig. 4. Instead, for every tested number of mixtures GFB

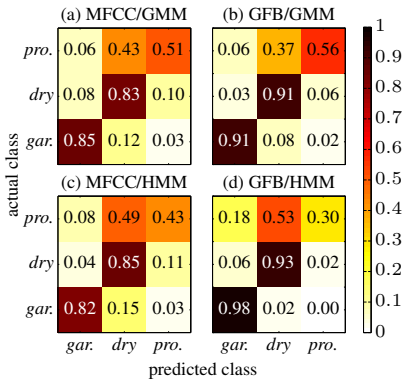

Fig. 5. Confusion matrices using MFCCs (panels (a) and (c)) and GFB features (panels (b) and (d)) for *dry cough*, *productive cough* and other sounds (*garbage*). In panels (a) and (b), every class is modeled by GMMs whereas in panels (c) and (d), classes *dry cough* and *productive cough* are modeled by three-states HMMs. Rows indicate the actual human labeled classes, columns the classifiers' predictions.

features yield better performance.

If merely discrimination of cough against other sounds is performed (cf. Fig. 3, left panel), hardly any differences in accuracy based on GMMs or HMMs are noticeable for each tested feature type. If *dry* and *productive coughs* are supposed to be differentiated as well, a benefit is achieved by applying GMMs (cf. Fig. 3, right panel). Though the correct recognition rate of *garbage* and *dry cough* increases by using HMMs in combination with GFB features, the accuracy for *productive cough* degrades in comparison to the usage of GMMs whereas confusions with *dry cough* and *garbage* increase as can bee seen in the confusion matrices depicted in Fig. 5.

6. CONCLUSION

In this contribution, cough classification with differentiation of productive and dry cough is conducted. GMMs and HMMs were tested as back-end classifiers in combination with MFCCs and GFB features.

We showed that for both back-end classifiers, GFB features yield higher accuracies than MFCCs, independent of the model parameterization. Furthermore, we demonstrated that GMMs yield higher accuracies than three-states HMMs if modeling *dry* and *productive cough*. This might be due to the annotations that do not distinguish between single coughs and cough series. In this case, an unstructured GMM seems beneficial over an HMM that explicitly attempts to model temporal configurations.

Hence, the accuracy for discrimination of garbage and cough based on GFB features in combination with GMMs is considerably high. However, distinction between *dry* and *productive cough* is still moderate.

References

[1] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, Aug. 2010, pp. 1267–1271.

[2] S. Goetze, J. Schröder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering (JCSE), SI on uHealthcare*, vol. 6, no. 1, pp. 40–50, March 2012.

[3] J. Schröder, N. Moritz, M. R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events'," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

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

[5] S. Matos, S. S. Birring, I. D. Pavord, and D. H. Evans, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1078–1083, 2006.

[6] A. A. Abaza, A. M. Mahmoud, J. B. Day, W. T. Goldsmith, A. A. Afshari, J. S. Reynolds, and D. G. Frazer, "Feature selection of voluntary cough patterns for detecting lung diseases," in *25th Southern Biomedical Engineering Conference*, Miami, Florida, USA, May 2009, pp. 323–328.

[7] V. Exadaktylos, M. Silva, S. Ferrari, M. Guarino, C. Taylor, J. Aerts, and D. Berckmans, "Time-series analysis for online recognition and localization of sick pig (sus scrofa) cough sounds," *Journal of the Acoustical Society of America (JASA)*, vol. 124, no. 6, pp. 3803–3809, 2008.

[8] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, and S. Kudoh, "Discrimination of productive and non-productive cough by sound analysis," *Internal Medicine*, vol. 37, no. 9, pp. 732–735, 1998.

[9] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, no. 6, pp. 2315–2331, Mar. 2000.

[10] L. M. Miller, M. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *Journal of Neurophysiology*, vol. 87, no. 1, pp. 516–527, Jan. 2002.

[11] J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich, "Spectrotemporal structure of receptive fields in areas ai and aaf of mouse auditory cortex," *Journal of Neurophysiology*, vol. 90, pp. 2660–2675, Jan. 2003.

[12] A. F. Meyer, J.-P. Diepenbrock, M. F. K. Happel, F. W. Ohl, and J. Anemüller, "Discriminative learning of receptive fields from responses to non-Gaussian stimulus ensembles," *PLoS ONE*, vol. 9, no. 4: e93062, Apr. 2014.

[13] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *INTERSPEECH*, Denver, Colorado, USA, Sep. 2002.

[14] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, May 2012.

[15] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, Aug. 2015, pp. 719–723.

[16] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal Gabor filterbank features for acoustic event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2198–2208, Dec. 2015.

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

[18] C. V. Cotton and D. P. W. Ellis, "Subband autocorrelation features for video soundtrack classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 8663–8666.

[19] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 493–497.

[20] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, USA, Oct. 2011, pp. 69–72.

[21] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 151–155.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 2006.