A TANDEM CONNECTIONIST MODEL USING COMBINATION OF MULTI-SCALE SPECTRO-TEMPORAL FEATURES FOR ACOUSTIC EVENT DETECTION

Miquel Espi[1]*, Masakiyo Fujimoto*[2]*, Daisuke Saito*[1]*, Nobutaka Ono*[3]*, Shigeki Sagayama*[1]

[1]Graduate School of Informations Science and Technology, University of Tokyo, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]National Institute of Informatics, Japan

ABSTRACT

Acoustic event detection systems supporting heterogeneous sets of events face the problem of having to characterize them when they have different acoustic properties (transient, stationary, both, etc.), observing this fact even within the acoustic event itself. Moreover, managing large feature vectors with features characterizing different properties of the signal is always difficult. This paper introduces the usage of spectro-temporal fluctuation features in a tandem connectionist approach, modified to generate posterior features separately for each fluctuation scale and then combine the streams to be fed to a classic GMM-HMM model. The experiments explore scale and event wise performance, as well as different stream combination methods, and show that the proposed method outperforms the GMM-HMM baseline as well as recent proposals in the CHIL 2007 evaluation campaign's related acoustic event detection tasks.

Index Terms— acoustic event detection, spectro-temporal fluctuation features, multi-stream combination, tandem connectionist

1. INTRODUCTION

Although most research in acoustic scene analysis has focused on speech, there are many other types of sounds, such as steps, door knocks, sneezes, breath, etc., that can provide valuable information in many applications, online and offline. The discovery and analysis of those sounds is referred to as acoustic event detection (AED). Its applications include audio recordings segmentation and surveillance, aliveness detection in elder care or automatic classification of social activities and contexts. A straightforward application of online AED would be systems that trigger other actions upon detecting a specific event such as notifying medical emergency services if breath cannot be detected anymore, or opening the door when somebody knocks, but also one can think of AED as previous stage to improve speech recognition systems by integrating this complementary information into adaptive noise reduction or feature enhancement systems. Furthermore, at the off-line side of AED there are many automated statistical applications that can improve our lives: e.g. counting the times one coughs or sneezes everyday, in combination with diseases diagnostics data, the relationship between both could be learned improving future diagnosis, among others.

Current most salient works in AED reflect the aim to bring most successful technologies of speech recognition to the field. Recently, CHIL evaluation framework 2007 [1] and its AED task has put effort to provide a common evaluation framework. In that evaluation campaign [2] achieved one of the best performances, using Adaboost-based feature selection with the Kullback-Leibler distance to measure the discriminant capability of each feature for each acoustic event and Hidden Markov Models (HMM). More recently, [3] present a two stage proposal: tandem connectionist stage, combining sequence modeling advantages with discriminative capabilities of Multi-Layered Perceptron (MLP) trained posteriors, and a re-scoring stage, in which similarly to speaker identification, boosts classification performance by adapting a Universal Background Model (UBM) Gaussian Mixture Model (GMM) to each class as the input to a Support Vector Machine (SVM) classifier providing confidence scores to each of the detected segments.

However, although these approaches include automated feature selection steps, little work has been done on including new features that would fit the targeted acoustic events. Another problem is that up to date no one from the AED field has explored different ways to integrate features characterizing different properties of the signal. More specifically, acoustic events have very different properties and it is reasonable to assume that the same set of features will not perform equally well in all cases — e.g. some events are more transient, some more stationary, some, like music, have components of both kinds. Therefore, there is a need to find features that would fit better the different kinds of acoustic events, and a way to integrate them in a consistent manner.

In this way, in [4] we introduced the usage of speech-specific spectral fluctuation filters [5] in place to obtain noise-wise robust performance of voice activity detection systems. In that case, the spectro-temporal parameters were set so that the resulting signal was to fit that of speech. Here, we propose the usage of spectral fluctuation related features characterizing the different sets of acoustic events such as the ones that are more transient, the ones that are more stationary, etc. In this paper we refer to these as fluctuation scales. Such a decomposition unfolds mixed information from the signal, generating a significantly large feature vector. That is the reason why we need to integrate that in a proper way so that we can take advantage of each of the decomposed components. Tandem connectionist models [6] combine the advantages of discriminative and generative models. In this work, we use multi-stream tandem modeling to achieve a discrminatively weighted integration of features from different fluctuation scales. For this purpose, we propose replacing traditional early integration scheme with a late integration scheme, previously proposed in [7] for integration of discriminative and generative evidence models. This is done by placing a group of posterior estimators (one for each of the fluctuation scale features), instead of a single one, and introducing a multi-stream integration step to integrate those posteriors, dealing better with a model that has features characterizing different properties.

The outline of this paper is as follows: Section 2 describes the spectral fluctuation related acoustic features included in the model and posterior features modeling, Section 3 summarizes the proposed stream integration and full models, Section 4 contains the evaluation experiments we performed and discusses the results, concluding the work in Section 5.

2. FEATURE EXTRACTION

2.1. Acoustic features to characterize acoustic events

Sounds found in an acoustic scene can be characterized at many domains: energy, spectrum, temporal variation of the spectrum, etcetera. However, such properties of speech fail to differentiate between events since even in spectrum-domain, noise and other acoustic components are mixed in the same signal. Two-stage Harmonic Percussive Sound Separation (HPSS) [5] enables to decompose a signal spectrum by its spectro-temporal smoothness, that is, by differentiating sounds by its smoothness in time and smoothness in spectrum at the same time, and it is based in the algorithm by the same name HPSS introduced in [8]. As it can be observed in Fig. 1, the HPSS algorithm is able to characterize mixed signals by their temporal and spectrum smoothness (represented in the horizontal and vertical axes of the graphs, respectively). Although this algorithm was originally used in music processing, the harmonic and percussive components obtained after the decomposition can be generalized as the components smooth-in-time and smooth-in-frequency, respectively. Further explanations on how the algorithm achieves the decomposition in an expectations-maximization fashion, can be found in [8].

The decomposition is parameterized by defining the Short-Time Fourier Transform (STFT) analysis window width which has a trade-off between its width in time and frequency. Two-stage HPSS takes this decomposition property to extract the parts from the signal that fluctuate in a desired manner (i.e. separates from the signal the components that do not fluctuate as desired), by applying the HPSS algorithm twice to filter with upper and lower smoothness boundaries. Such property was used in [4], and here the same approach is applied but this time taking into account also the rejected components during the two stages, given they are more prone to characterize signals that are transient, or prominently stationary better than a raw spectrum. As a result we obtain three spectrograms: one containing components rather stationary (H-component), a second containing more transient components (P-component), and a third containing the intermediate part (V-component) that has been previously used in voice activity detection research [4]. These three spectrograms, together with the original are the four fluctuation scales we refer to when we talk about stream combination, and they have been characterized by extracting their Mel-Frequency Cepstrum Coefficients (MFCC), together with their first and second derivatives, as acoustic features.

2.2. Posterior features

Following the tandem connectionist approach proposed for speech recognition in [6], much discussion has arisen mostly focused on justifying the advantages and disadvantages of such an approach. In this way, the core of the tandem connectionist approach is the usage of the posteriors probabilities (for each phone in the case of speech recognition) as features, namely posterior features, to be fed to a generative model such as a conventional GMM-HMM instead of using the acoustic features themselves. However, although several ways of obtaining those posterior features have been proposed, discriminatively trained neural networks such as MLP have provided the best results so far. Looking deeper, the combination of the discriminative (MLP) and generative (GMM) models provides: first, enhancement of the variation and minimize irrelevant detail in feature space due to the fact that MLPs modeling focuses on small patches on the boundaries between acoustic events, magnifying the boundary spaces; and second, combines this with a *pseudo language model* provided by the GMM-HMM. This is, it allows modeling of the feature stream as a sequence, which is similar to some extent to how languages provide a priori information.

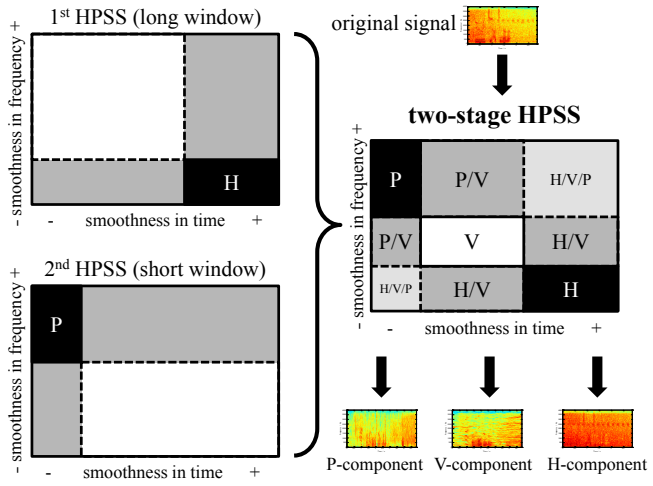

Fig. 1. Two-stage HPSS components location scheme in a spectro-temporal smoothness representation: H-component for stationary signals, P-component for transient signals, and V-component for the intermediate component.

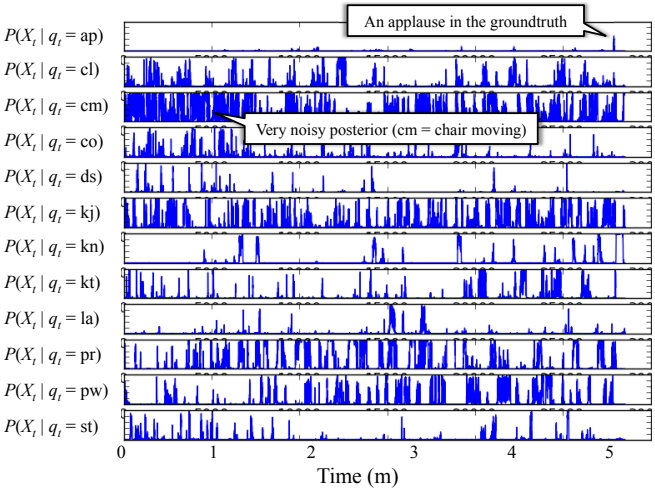

Fig. 2. Example of the posteriors of each event ($P(X_t|q_t =$[event]$)$) obtained from a sample recording session using the trained MLPs.

An example of posterior features can be observed in Fig. 2. It depicts how MLP posterior provide very clear information in some cases such as the *applause* posterior, and also noisy posteriors as it can be seen in the posterior for *chair moving*. Here, the aim of using these posteriors as features for a likelihood-based model is to leverage also information from other posteriors and combine them.

3. MULTI-STREAM COMBINATION

Traditionally, combining feature streams characterizing different properties of the signal is performed by using a large feature vector or by applying some dimensionality reduction techniques such as Principal Component Analysis (PCA). Here we propose the integration of the stream combination in the tandem connectionist model by modifying the posterior features stage into a set of hierarchical MLPs grouped by features of the same fluctuation scale, and then

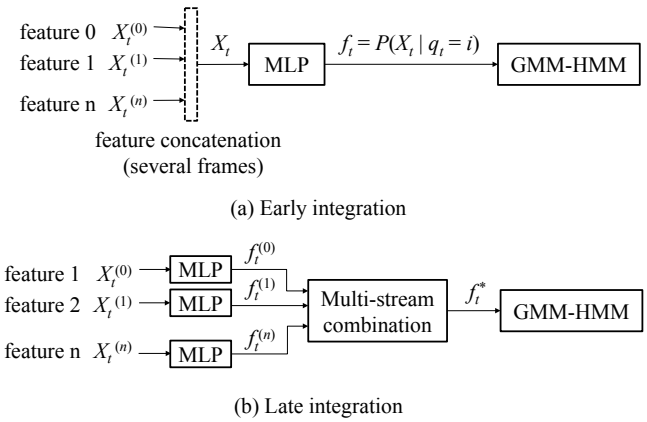

Fig. 3. Integration schemes for combining multiple domain features: (a) Early integration, and (b) Late integration.

combine the resulting posteriors streams by acoustic event. This integration schemes are based on the concepts introduced in [7] by taking the early and later integration models to compare their performance, but used here to combine different acoustic natures. The two models are described in Fig. 3.

Early integration Fig. 3 (a) refers to the common approach in tandem connectionist architectures. In this combinations scheme the acoustic features are concatenated no matter what signal property they belong to they are from, and fed to a MLP to estimate the posteriors (f_t) which are ultimately used as features to classifier likelihood estimation, and path decoding (GMM-HMM in this case). On the other side, late integration refers to the integration of the property grouped features once the posteriors have been obtained by means of a set of hierarchical posterior estimators (e.g. MLPs). These are integrated later on using sum, weighted sum, product, or even PCA, to integrate the posterior feature, that will be again fed to a GMM-HMM model. The full model flow is shown in Fig. 3 (b), and is a modified version of the traditional tandem connectionist approach in which the knowledge of the signal property characterized by each feature is used to group features, and obtain multiple posteriors ($f_t^{(i)}$) that are later on combined into (f_t^*).

Having extensive literature in stream combination, here we have investigated the usage of sum [9, 10], and PCA, to evaluate and compare their performances in late integration. The performance of early integration has also been tested for comparison purposes, as well as traditional GMM-HMM.

4. ACOUSTIC EVENTS RECOGNIZER EVALUATION

This section describes a first experiment consisting in AED in scenarios where acoustic events happens without overlapping with other acoustic events. The target of this experiment is to analyze in detail the behavior of each of the proposed features, as well as different stream integrations models, and discuss their advantages.

The task consisted in performing AED in acoustic scenes with isolated acoustic event, i.e. only overlapped with the environmental noise. The recordings belong to the FBK-irst AED development set contained in CHIL evaluation framework 2007 [1] (and are also publicly available through ELRA repository[1]) and the description of the room can be found in the referred paper. The acoustic events included are: knock on a door or table (kn); door slam, on open or close (ds); steps (st); chair moving (cm), spoon clings or cup jingle (cl); paper wrapping (pw); key jingle (kj); keyboard typing (kt); phone ringing/music (pr); applause (ap); cough (co); laugh (la); and unknown (no). Although the dataset features recordings from several distant microphone arrays, and microphones on the table. In this experiment we used one of the microphones on the table as only acoustic input. From the dataset we took 6 sessions of the 9 available to train the MLPs, and the posterior features GMMs, and the rest (3 sessions) for the tests.

Table 1. AED-accuracy for each of the models in isolated AED

Model	AED-ACC
GMM-HMM	18.60%
MLP-GMM-HMM (early)	26.32%
MLP-GMM-HMM (late-PCA)	20.31%
MLP-GMM-HMM (late-sum)	**49.73%**

The data was down-sampled from its original 44.1 kHz sampling rate to 16 kHz, and MFCCs (12 coefficients and energy) were extracted from the 3 resulting components of the two-stage HPSS, and from the original signal, together with their first and second derivatives. Resulting into a total of 156 dimensions feature vector ($4\times13\times3$) in the case of early integration scheme, and 4 sets of 39 dimensions features vectors in the case of late integration. MLPs have been trained and developed using QuickNet[2] which provides efficient tools to manage MLPs. In this experiments MLPs input included a context of 9 frames (4 frames around the current frame) to model the trajectory of the features as well. The MLP in the early integration MLP-HMM model has way more inputs than the MLPs in the later integration case. Therefore, to ensure a fairness in the comparison, the early integration MLP (512 nodes) has more nodes in the hidden layer than the MLPs (128 nodes) in the late integration scheme, however, it is difficult to say how far this is affecting the performance.

As defined in CHIL 2007 evaluation framework, we consider an acoustic event has been correctly detected when either the central frame of the detected event period falls within the event in the groundtruth reference, or when the central frame of an event in the groundtruth falls within a detected event. This is because, unlike in speech detection where we require the full speech utterance to perform later processing such as speech recognition, AED systems treat events as timestamps (e.g. the target is to know if somebody knocked on the door or not, or count the number time one sneezes, etcetera). With this we evaluate the performance of the experiments at 2 levels: frame-wise, by obtaining the confusion matrix between acoustic events, and the hit rate of each of them; and segment-wise, where we use the AED-accuracy measure [1] which accounts for the harmonic mean between precision and recall, these being the fraction of the detected acoustic events that are in the reference and the fraction of the reference acoustic events that have been detected.

For the sake of performance comparison between multi-stream combination options, 3 different models have been tested:

- GMM-HMM: a traditional GMM-HMM ergodic model including a garbage state.
- MLP-GMM-HMM (early): tandem connectionist model with early integration of different fluctuation scale features and posterior features obtained from the MLP.
- MLP-GMM-HMM (late): tandem connectionist model with late integration (ranging within sum, product) of the posterior obtained with MLPs.

[1] http://catalog.elra.info/product_info.php?products_id=1093

[2] http://www.icsi.berkeley.edu/Speech/qn.html

Table 2. Confusion matrix for isolated AED: percentage between the detected events and the groundtruth events.

Groundtruth	**Detected event**											
	ap	cl	cm	co	ds	kj	kn	kt	la	pr	pw	st
applause (ap)	**22.5%**	3.2%	1.6%	6.4%	0.0%	25.8%	0.0%	3.2%	0.0%	0.0%	35.4%	1.6%
spoon cling (cl)	0.0%	**36.0%**	3.1%	0.4%	0.0%	22.1%	0.4%	29.9%	2.1%	2.5%	0.4%	2.7%
chair moving (cm)	0.0%	21.7%	**39.6%**	2.7%	8.1%	1.9%	1.0%	3.2%	2.1%	5.4%	2.7%	30.7%
cough (co)	1.3%	12.1%	18.7%	**34.5%**	8.5%	5.5%	0.0%	3.2%	1.9%	8.5%	3.2%	1.9%
door slam (ds)	0.0%	0.5%	20.5%	1.7%	**49.4%**	5.8%	1.1%	0.5%	0.0%	0.0%	0.5%	24.7%
key jingle (kj)	0.3%	16.6%	7.3%	1.5%	0.0%	**34.0%**	1.7%	17.7%	0.0%	6.4%	11.5%	2.5%
knock (kn)	0.0%	0.0%	5.9%	0.0%	1.7%	4.2%	**53.8%**	1.7%	4.2%	9.4%	0.0%	18.8%
keyboard typing (kt)	0.0%	12.4%	1.9%	0.0%	0.6%	26.1%	0.6%	**41.4%**	0.3%	2.4%	6.2%	7.6%
laugh (la)	0.0%	8.9%	16.7%	6.6%	6.2%	3.7%	2.9%	4.8%	**23.7%**	19.7%	0.7%	9.2%
phone/music (pr)	0.0%	19.0%	8.8%	3.7%	1.1%	2.3%	1.3%	9.0%	2.7%	**46.7%**	0.9%	3.9%
paper wrapping (pw)	0.8%	4.8%	6.0%	0.0%	0.0%	23.6%	0.0%	21.6%	0.4%	3.2%	**38.0%**	1.6%
steps (st)	0.0%	1.8%	23.0%	0.0%	10.9%	1.3%	2.9%	5.3%	0.5%	2.1%	0.0%	**51.7%**

Table 1 summarizes the AED-accuracy results of each of the systems tested and it can be observed that significant improvement obtained from using the late integration model. It is also significant the fact that also the traditional tandem connectionist model out-performs the the GMM-HMM as it was learned in [3]. Additionally, the confusion matrix between the actual and detected acoustic events can be found in Table 2. Here it can be observed which acoustic events are more prone to be confused with others, and to which extend, reflecting frame-wise performance. This defeats the meaning of acoustic event to some extent, since the main target of AED is to obtain timestamps, and frame level performance is not a main target. However, it is interesting to observe the low rates of confusion obtained for applause (ap), laugh (la). Other than that, the results are all coherent with the accuracy results. We also performed False Alarm and False Rejection Rates (FAR, and FRR, respectively) which revealed that also these measure improve from using tandem connectionist model with early integration (FAR of 8.75%, and FRR of 62.17%), to late integration with sum stream combination (FAR of 1.75%, and FRR of 40.83%).

5. CONCLUSION

As shown above, in the experiment we have compared the performance in AED between traditional GMM-HMM, tandem connectionist with early integration, and tandem connectionist with late integration schemes, in AED of isolated acoustic events, concluding that the usage of posteriors of each acoustic event as features fed for a GMM-HMM model performs better than using acoustic features directly in a generative model. Moreover, we observed that there is a clear benefit in using late feature integration both in complexity and performance. These conclusions lead to think that exploiting the modified schemes of the tandem connectionist model can provide an advantage when combining features from different fluctuation scales in the same model. We also assume that more extensive error analysis has to be done in terms of variating the number of hidden nodes in the MLPs, among other parameters to obtain more accurate conclusions. However, this study might tackle some questions to be solved in future works.

Additionally, yet another question remains to be answered regarding the analogy with speech recognition. Such analogy approach has provided great results in other fields (e.g. music information retrieval), however, acoustic scene analysis might require a different paradigm to exploit properly most successful technologies of speech recognition.

6. ACKNOWLEDGMENTS

Parts of this research was conducted at NTT Communication Science Laboratories, and we would like to thank Dr. Shinji Watanabe and Dr. Yotaro Kubo, members of this laboratory, for their suggestions and valuable discussion during the process of this research.

7. REFERENCES

[1] A. Temko, "CLEAR 2007 AED evaluation plan and workshop," *http://isl.ira.uka.de/clear07*, 2007.

[2] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T.S. Huang, "HMM-based acoustic event detection with adaboost feature selection," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., vol. 4625 of *Lecture Notes in Computer Science*, pp. 345–353. Springer Berlin / Heidelberg, 2008.

[3] X. Zhuang, Z. Zhou, A. Hasegawa-Johnson, and T.S. Huang, "Real-world acoustic event detection," *Pattern recognition Letters*, vol. 31, pp. 1543–1551, 2010.

[4] M. Espi, S. Miyabe, T. Nishimoto, N. Ono, and S. Sagayama, "Using spectral fluctuation of speech in multi-feature hmm-based voice activity detection," in *Proc. of INTERSPEECH*, 2011, pp. 2613–2616.

[5] H. Tachibana, N. Ono, and S. Sagayama, "Vocal sound suppression in monaural audio signals by multi-stage harmonic-percussive sound separation (HPSS)," in *Proc. of ASJ Spring Meeting*, 2009.

[6] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 3, pp. 1635–1638.

[7] J. Pinto and H. Hermansky, "Combining evidence of from a generative and discriminative model in phoneme recognition," in *Proc. of Interspeech*, 2008, pp. 2414–2417.

[8] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. of EUSIPCO*, 2008, pp. 1–4.

[9] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[10] L.I. Kuncheva, "A theoretical study on six classifier fusion strategies," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 281–286, 2002.