MULTI-SOURCE POSTERIORS FOR SPEECH ACTIVITY DETECTION ON PUBLIC TALKS

Marc Ferràs and Hervé Bourlard

Idiap Research Institute, CH-1920 Martigny, Switzerland

marc.ferras@idiap.ch, herve.bourlard@idiap.ch

ABSTRACT

Speech activity detection (SAD) is a conceptually simple task that still poses serious challenges for speech processing in a large variety of scenarios. Current energy-based and model-based approaches tend to directly segment speech and non-speech classes, but are not robust enough to non-stationary noise. In this paper, we use a multi-source activity detection (MSAD) approach to SAD by finding the activity levels of speech and a set of non-speech acoustic sources. Public talks such as TED involve a large variety of non-speech audio that is difficult to handle with standard SAD systems. We evaluate the effect of using either the proposed MSAD system versus a tailored version of the popular SHOUT SAD system. We evaluate our approach on a subset of the TED data to show the effectiveness of the technique, with and without a sparsity constraint on the vector of acoustic source activities.

***Index Terms*—** multi source activity detection, speech activity detection, sparse constraint, speaker retrieval, speaker diarization

1. INTRODUCTION

The recording of public talks in symposiums, conferences and lectures typically entails high quality recording equipment set up by professionals. However, the room acoustics in these scenarios tends to be fairly reverberant since, even if close-talking and lapel microphones are used to capture the speech, a sense of the acoustic background is typically seeked for. Besides the main speaker, or few speakers at most in multi-speaker talks and interviews, all sorts of feedback from the audience such as applause, laugh, woos, and whistles are also recorded. Rehearsed public talks such as those from Technology, Entertainment, Design (TED) [1] include recorded and live music as well as adverts resulting in a wide variety of acoustic scenes.

Technologies such as speech recognition and speaker diarization can output cues that are valuable for the search and retrieval of semantically relevant information in a multimedia archive. Algorithms aiming at automatically processing the type of data described above typically assume knowledge of speech regions in the audio, found using a Speech Activity Detection (SAD) system. Most SAD algorithms focus on quiet acoustic environments, sometimes involving some stationary noise, that shape the task as a two-class classification problem. Energy-based approaches tend to generalize well in quiet conditions, but fail to properly handle noise as non-speech. Preprocessing the audio using noise reduction techniques such as the Wiener filter, or Short-Time Speech Amplitude (STSA) or Log Spectral Amplitude (LSA) estimators, can make an energy-based SAD effectively handle stationary noise, although probably lacking robustness when it comes to non-stationary noise and more structured audio such as music. Model-based approaches use the speech and noise structure to determine whether speech or non-speech is present based on models trained on speech and non-speech audio. The popular SHOUT [2] package, designed for broadcast news data, uses an iterative segmentation/training algorithm to train speech, silence and sound models that are then used to segment the audio.

When non-speech audio is heterogeneous it seems natural to use more than one model to capture the variability of the different non-speech sources more precisely. In this paper, we approach speech activity detection as a multi-class problem, where one class is speech and the rest are different types of non-speech. Furthermore, we assume that more than one of these classes can be active simultaneously by using an activity level for each of them, a posterior probability indeed. Since very few of the available acoustic sources are expected to be active at a time, we also use a sparsity constraint to estimate the vector of class posterior probabilities. The framework is based on the work presented in [3] on speaker detection where the activity levels are the weights of a mixture model. These weights are estimated using an EM algorithm and the objective function can be augmented with a sparsity constraint. The current work was encouraged by the positive results obtained in this study.

Section 2 describes the theoretical framework, underlying models and estimation procedures of the sparse and non-

This work was supported by the European Union under the FP7 Integrated Project inEvent (Accessing Dynamic Networked Multimedia Events), grant agreement 287872. The authors gratefully thank the EU for their financial support and all project partners for a fruitful collaboration.

sparse algorithms. Section 3 gives details about the training and detection setups of the multi-source activity detection algorithm. Section 4 provides and discusses the speech activity detection error results obtained for a set of TED talks and Section 5 draws some conclusions out of this work.

2. MULTI-SOURCE ACTIVITY DETECTION

The multi-source activity detection (MSAD) system uses a latent variable approach to estimate the activity level of each acoustic class. It is based on the work presented in [3], where the activity levels of several speakers in a mixture of speakers are detected in the presence of convolutive distortion. This framework assumes that feature vectors $\mathbf{x}_t$ at time t originate from a mixture of known mixture models with pdf

$$p(\mathbf{x}_t, \boldsymbol{\alpha}) = \sum_{m=1}^{M} \alpha_m p(\mathbf{x}_t; \lambda_m) \quad \text{with} \quad \boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}, \tag{1}$$

where α_m indicates the activity of class m with $\sum_{m=1}^{M} \alpha_m = 1$, and $p(\mathbf{x}_t; \lambda_m)$ is the pdf of class m estimated via Gaussian Mixture Models (GMM) or Student-t Mixture Models (TMM) [4] in this work. The hypothesis motivating such model is that the non-linearities in the feature space can be captured by linear relations in the probability space. A close look into (1) reveals that the proposed model is adequate for the single source case. However, a weighted sum of pdfs represents a multiple choice scenario rather than the non-linearities found in spectral envelope features of overlapping speech. It still provides a mechanism to assign probabilities to each class while definitely simplifying the estimation procedure of the activity levels.

The activity levels of the acoustic sources, i.e. the vector $\boldsymbol{\alpha}$, are estimated on a window of $2T + 1$ feature vectors. In this case, the total likelihood for a given window becomes

$$p(\mathbf{X}, \boldsymbol{\alpha}) = \prod_{t-T}^{t+T} p(\mathbf{x}_t, \boldsymbol{\alpha}) \tag{2}$$

if we assume independence amongst frames. As we describe below, an EM algorithm can be used to find the $\boldsymbol{\alpha}$ that maximizes the total likelihood function.

2.1. Non-sparse $\boldsymbol{\alpha}$

Equation 1 has the form of a GMM, with the Gaussian pdfs for each mixture being replaced by generic distributions $p(\mathbf{x}_t; \lambda_m)$. Estimation of the α_m weights is hence analogous to the maximum likelihood estimation of Gaussian weights of GMM [5]. The method of Langrange multipliers is used to maximize (2) with the constraint that the weights sum up to one. An Expectation-Maximization (EM) algorithm provides weight estimates by alternating the computation of the class posterior probabilities by

$$\gamma_{mt} = \frac{\alpha_m p(\mathbf{x}_t, \lambda_m)}{\sum_{l=1}^{M} \alpha_l p(\mathbf{x}_t, \lambda_l)} \tag{3}$$

and then iteratively updating the weights according to

$$\alpha_m = \frac{1}{T} \sum_{t=1}^{T} \gamma_{mt} \quad . \tag{4}$$

These update equations are the same as used for mixture weight training of GMM. The algorithm is guaranteed to converge to local maximum likelihood weight estimates.

2.2. Sparse $\boldsymbol{\alpha}$

In audio recordings, acoustic sources are rarely active at the same time, especially in the case of communication. In a public talk, the speaker, the audience and also the technical staff make it such that the number of acoustic sources are few enough to be distinguishable and understandable. This can be modeled by a sparsity constraint on the components of $\boldsymbol{\alpha}$. As proposed in [3] the total likelihood function of (2) can be regularised by adding a term measuring the sparsity of $\boldsymbol{\alpha}$. We use Hoyer's sparsity measure [6]

$$H(\boldsymbol{\alpha}) = \frac{1}{\sqrt{M} - 1} \left(\sqrt{M} - \frac{||\boldsymbol{\alpha}||_1}{||\boldsymbol{\alpha}||_2} \right) \tag{5}$$

for this purpose. This measure yields a value equal to 1 if and only if one component α_m is not 0, zero if all the components are equal. The new objective function becomes a linear combination of the total likelihood and sparsity terms as

$$\boldsymbol{\alpha}^* = \arg\max_{\alpha} \left(\theta p(\mathbf{X}, \boldsymbol{\alpha}) + (1 - \theta) H(\boldsymbol{\alpha}) \right) \tag{6}$$

subject to $\sum_{m=1}^{M} \alpha_m = 1$. Note that whenever $\theta < 1$ the estimated weights will be suboptimal in likelihood terms. The range $0 \leq \theta \leq 1$ sets a trade-off from pure likelihood to pure sparsity.

As in the non-sparse case, the EM update equations are found by setting the new Lagrangian to zero and then finding its derivatives. However, in the sparse case, the maximization step has no closed-form solution and the non-linear set of M equations as

$$\begin{gathered} \theta(\sqrt{M} - 1)\, ||\boldsymbol{\alpha}||_2^3 \left(\sum_{t=1}^{T} \gamma_{mt} - T\alpha_m \right) \\ + (1 - \theta)\alpha_m \left(\alpha_m - ||\boldsymbol{\alpha}||_2^2 \right) = 0 \end{gathered} \tag{7}$$

on M variables is to be solved. Using the non-sparse estimates of $\boldsymbol{\alpha}$ as initial conditions we solve such system of equations using the Levenverg-Marquardt algorithm.

3. MSAD SYSTEM SETUP

The following sections give details about the model training step and the event detection step of the MSAD system.

3.1. Model training

The MSAD system assumes statistical knowledge about the acoustic sources to detect, which translates into having a way to compute $p(\mathbf{x}_t; \lambda_m)$, for all m. In this work, these are either Gaussian or Student-t Mixture Models (GMM,TMM) trained using a small set of the TED [1] corpus, i.e. public talks and lectures respectively, and a subset of the Magnatune [7] corpus, a music database featuring a permissive license. The details for the training data are shown in Table 1. Targeting public talks we find it interesting to focus on six acoustic classes, namely speech, silence, laugh, applause, music and TED music. The latter models the TED head/tail music that would eventually allow us to safely ignore audio before and after it. The labels for each class were obtained manually, paying special attention to use audio with a single acoustic class only. All the acoustic classes except music were trained on TED data, i.e. in-domain data. For the music class, a selection of songs covering a variety of styles were taken from the Magnatune database to train the music model.

Class	Sources	Recordings	Length(min)
Speech	TED	12	16
Silence	TED	12	5
Laugh	TED	12	2
Applause	TED	12	3
TED Music	TED	12	4
Music	Magnatune	51	57

Table 1. Data sources, number of recordings and audio length used for training the acoustic event models.

We trained a 64-mixture GMM using 19 MFCC as features and 5 iterations of maximum likelihood estimation. We also trained TMM, whose mixtures have heavier tails than Gaussians, becoming less sensitive to outliers during training. Since the TMM parameteres tend to converge rather slowly, the TMM were bootstrapped from the GMM and setting the initial degrees of freedom to 20 for each mixture, resulting in peaky pdfs. All parameters, i.e. weights, means, variances and degrees of freedom were then reestimated using an EM algorithm [4] with 10 iterations.

3.2. Activity Detection

Once the models are trained, the algorithms described in Sections 2.1 and 2.2 are used to determine the activation level of each class on a sliding window of 300ms and a window shift of 10ms. These values result in a satisfactory trade-off between speed and performance as informally seen in some preliminary tests. We also apply a windowing function onto the likelihood scores used to compute the class posterior probabilities (3) so that the center frames are given larger weights compared to those at the boundaries. Sharper decision edges over time as well as more stable steady states were informally observed using this windowing function.

4. EXPERIMENTS AND RESULTS

We ran a set of experiments comparing the SHOUT [2] speech vs. nonspeech detection system with the proposed MSAD approach. There are a number of differences in both approaches that are worth mentioning. In this work, SHOUT was trained to detect three acoustic classes, namely speech, silence and sound using exactly the same data used for the MSAD system. MSAD uses 6 acoustic models, speech, silence, laugh, applause, music and TED music instead. SHOUT uses the Viterbi algorithm during decoding on a frame-by-frame, i.e. every 10ms, basis whereas the MSAD approach uses no decoding outputing class posteriors every 100ms. Both algorithms use 19 MFCC features plsu log-energy together with their corresponding delta and double delta coefficients.

In the International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns manually annotated speech activity labels are provided. We use the 2010 development and the 2010, 2011 and 2012 evaluation sets together as the evaluation data set for the speech activity detection experiments. The compound data set involves 31 TED talks with around 7 hours of audio manually annotated with speech activity labels. We use the average miss, false alarm and total errors as performance measure. The detected class is found by maximum likelihood decoding for SHOUT and by taking the class with maximum posterior probability for the MSAD system. A smoothing time of 1.5s was applied on the output of both systems. No collar was used during scoring.

Table 2 gives the results for the SHOUT and several setups of the MSAD system. The MSAD system outperforms the SHOUT system by 7% to 9% in terms of relative total error. The best absolute error rate is 4.9%, meaning that around 95.1% of the time the MSAD system gives the right specch/nonspeech label. The MSAD system gives miss and false alarm errors in the same range whereas a large proportion of the errors are false alarms for the SHOUT system. Note that the SHOUT system outputs the class activities that maximize the likelihood over the whole recording whereas the MSAD system uses the maximum posterior probability at each time frame, that is, locally optimized. Neither of these systems has an easy way of tuning the operating point, although an insertion penalty could be eventually used during Viterbi decoding. We did not explore this possibility in this paper.

Regarding the MSAD system, error rates stay stable across all setups. There is no clear difference in using Gaussian or Student-t Mixture Models. The additional complexity of TMM is a retraining pass of all mean, variance, weight and degrees of freedom parameters whereas the likelihood computation cost is still comparable to that of Gaussian components.

The MSAD setups using the sparsity constraint ($\theta = 5e-3$) obtain marginal gains only when using TMM and do not bring any improvement otherwise. We have observed that the non-sparse MSAD system ($\theta = 1$) outputs precise class activities almost all the time. This suggests that the acoustic models are pure enough for the posteriors to be clean whenever only one source is active, something expectable after manual annotation of the training data. In practice, the sparsity constraint has a large effect on the activity contours over time as shown in Figure 1, although differences against the non-sparse system are small if we take the winning class for each frame into account. On the other hand, one or two active components in a vector of six may not be necessarily considered as sparse, especially when we compare to other studies such as [3] or the vast majority of work on sparse linear systems that consider thousands of components.

SAD	θ	Miss(%)	FA(%)	Error(%)
SHOUT	-	1.4%	4.0%	5.4%
MSAD-GMM	1	2.1%	2.8%	4.9%
MSAD-GMM	5e-3	2.2%	2.8%	5.0%
MSAD-TMM	1	2.2%	2.8%	5.0%
MSAD-TMM	5e-3	2.2%	2.7%	4.9%

Table 2. Miss, false-alarm and total errors of the SHOUT speech/non-speech detection system and the proposed multi-source activity detection system.

5. CONCLUSION

We showed the effectivity of the proposed multi-source activity detection (MSAD) system in the context of public talks such as those from TED. In such scenario, this system is able to properly deal with non-stationary noise and sound from the acoustic environment, the speakers and the audience for the purpose of speech activity detection. The MSAD obtained relative gains in speech detection error in the range 7% to 9% versus the SHOUT speech/non-speech segmentation system that uses a speech, silence and sound models. The different variants of the MSAD algorithm did not result in significant differences in error rate, whether Gaussian mixtures models or Student-t mixture models were used. We also showed the use of a sparsity constraint on the vector of activities that improved the visualization aspects of the activity contours over time while it turned out not to be effective in error rate terms.

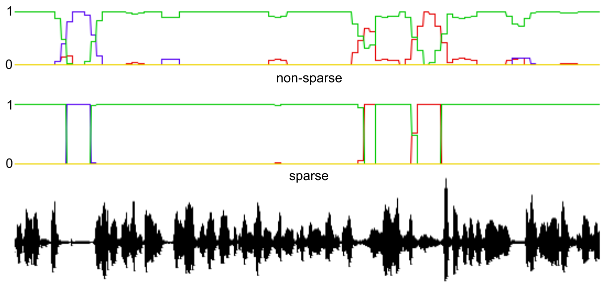

Fig. 1. Activity contours of non-sparse ($\theta = 1$, top) and sparse ($\theta = 5e-3$, middle) MSAD system outputs over time. The waveform is shown below for reference.

6. REFERENCES

[1] "Technology, entertainment design," `http://www.ted.com`, Accessed October 23rd, 2013.

[2] M.Huijbregts and F. de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[3] H. Sundar, T.V. Sreenivas, and W. Kellermann, "Identification of Active Source in Convolutive Mixtures using Knwon Source Models," *IEEE Signal Processing Letters*, vol. 20, pp. 153–156, 2013.

[4] D. Gerodiannis, C. Nikou, and A. Likas, "The Mixtures of Student's t-distributions as a Robust Framework for Rigid Registration," *Journal of Image and Vision Computing*, vol. 27, pp. 1285–1294, 2009.

[5] J. Bilmes, "A gentle tutorial on the EM algorithm including gaussian mixtures and baum-welch," *Technical Report TR-97-021, International Computer Science Institute, Berkeley, CA*, 1997.

[6] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[7] "Magnatune," `http://www.magnatune.com`, Accessed October 23rd, 2013.