INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Indoor/Outdoor Audio Classification using Foreground Speech Segmentation

Banriskhem K. Khonglah[1], *K. T. Deepak* [2] *and S. R. Mahadeva Prasanna*[1]

[1]Department of Electronics and Electrical Engineering
Indian Institute of Technology (IIT) Guwahati, Guwahati-781039, India
[2]Department of Electronics and Communication Engineering
Indian Institute of Information Technology (IIIT) Dharwad
{banriskhem, prasanna}@iitg.ernet.in, deepakkt@iiitdwd.ac.in

Abstract

The task of indoor/ outdoor audio classification using foreground speech segmentation is attempted in this work. Foreground speech segmentation is the use of features to segment between foreground speech and background interfering sources like noise. Initially, the foreground and background segments are obtained from foreground speech segmentation by using the normalized autocorrelation peak strength (NAPS) of the zero frequency filtered signal (ZFFS) as a feature. The background segments are then considered for determining whether a particular segment is an indoor or outdoor audio sample. The mel frequency cepstral coefficients are obtained from the background segments of both the indoor and outdoor audio samples and are used to train the Support Vector Machine (SVM) classifier. The use of foreground speech segmentation gives a promising performance for the indoor/ outdoor audio classification task.

Index Terms: indoor, outdoor, foreground speech segmentation, SVM

1. Introduction

The broadcast audio processing and transcription is a challenging task considering the large variabilities present in the data. There are many studies that attempt to process broadcast audio [1–11]. The steps followed in most of the previous works consists of a preprocessing step followed by the speech recognition step. The preprocessing step consists of tasks such as speech/non-speech detection, gender detection and bandwidth detection [1, 3]. The speech recognition step involves training the models using mel frequency cepstral coefficients (MFCCs) as features [1, 2, 12]. Alternatively, this work attempts to look at the broadcast audio processing in a different manner.

The data present in broadcast audio generally consists of either the anchor's speech or the reporter's speech. In most of the cases, anchor's stay indoor while reporters speak from outdoor environments. In such scenario, the anchor's speech is relatively clean, while the speech recorded from reporters in outdoor environments contain background noise. The indoor speech is clearly audible to listeners and can be used for speech to text transcription. Due to the presence of high background noise, the outdoor speech may not be clear to listeners and further it may not be suitable for speech to text transcription. Hence, it is necessary to enhance the speech signal to make it suitable for listening and speech to text transcription. It is imperative to segment the recorded broadcast audio into indoor and outdoor before enhancing the audio segments. In this work a new approach is proposed to classify the audio segments into indoor and outdoor. The focus is mainly on pre-recorded audio signals and not on-line.

The indoor and outdoor audio samples being recorded in different scenarios will have different types of acoustic environments. For example, the indoor speech signal has the least interference from other acoustic sources, but the outdoor speech is affected by the presence of other acoustic sources. The speech signal from anchor or reporter speaking closer to microphone is termed as *foreground speech* and rest of the interfering acoustic sources are categorized as *background noise*. This difference in the background environments of the two segments enables to explore the classification task in terms of the foreground speech segmentation. The method for foreground speech segmentation is recently proposed in [13, 14]. There is a difference in signal characteristics due to the variations in the distance between foreground speaker to microphone compared to rest of the background sources. The method exploits this signal characteristics to segment the foreground speech from rest of the background noise. For the indoor and outdoor audio recordings, the desired speaker is closer to the microphone. However, the background noise due to the other interfering sources are recorded at a farther distance from the microphone. The background noise segments obtained complementary to foreground speech segments can be exploited to perform the indoor and outdoor classification task.

The foreground speech segmentation method segments a given audio sample into foreground and background regions using the normalized autocorrelation peak strength of the zero frequency filtered signal [13]. This method will be utilized to segment the foreground and background regions of the indoor and outdoor audio. The background segments can then be considered for the classification task. The mel frequency cepstral coefficients (MFCC) are extracted from the background segments and classified using the support vector machine (SVM) classifier. The rest of the work is organized as follows. The foreground speech segmentation along with the classification method are described in section 2. Section 3 describes the results and discussion and section 4 concludes the work.

2. Foreground Speech Segmentation using NAPS of ZFFS

The foreground speech segmentation consists of using a features derived from the zero frequency filtered signal (ZFFS) [15, 16]. The ZFFS is obtained as follows,

- Difference the speech signal $s[n]$

$$x[n] = s[n] - s[n-1] \tag{1}$$

- The differenced speech signal $x[n]$ is passed through a cascade of two ideal zero frequency (digital) resonators,

http://dx.doi.org/10.21437/Interspeech.2017-309

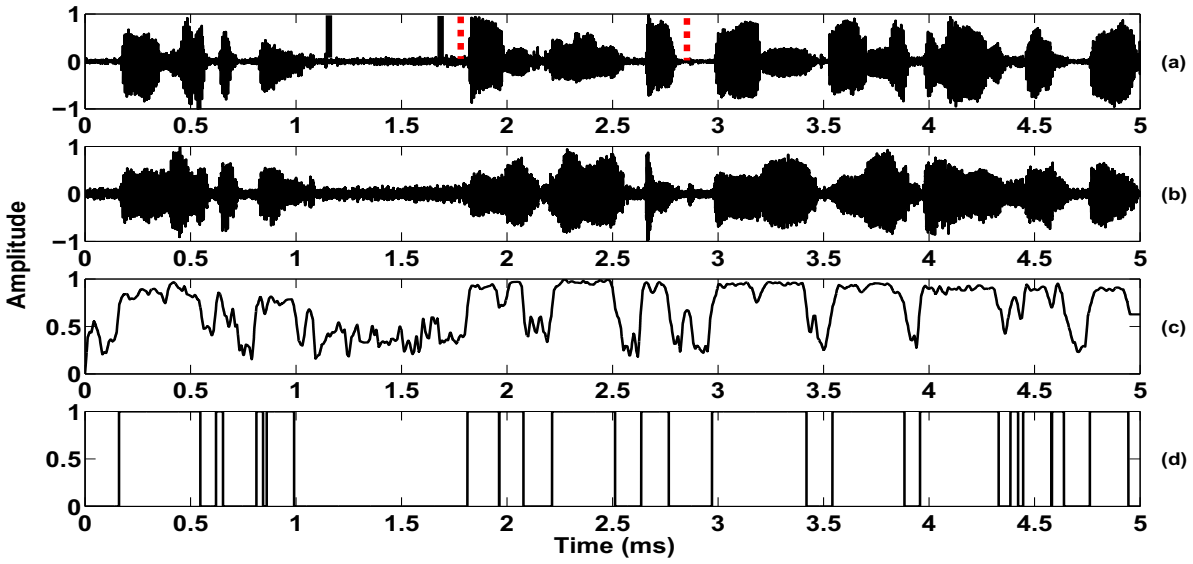

Figure 1: *Illustration of foreground speech segmentation on a sample of 5 s of outdoor speech. (a) Speech signal (b) ZFFS of speech (c) Normalized autocorrelation peak strength (NAPS) of ZFFS (d) Non-linearly mapped value of (c) in which the foreground regions are mapped to one and the background regions are mapped to zero. A sample of a foreground region is marked with the dotted red line and a background region is marked with the continuous black line*

i.e,

$$y[n] = -\sum_{k=1}^{4} a_k y[n-k] + x[n] \tag{2}$$

where a_1 = -4, a_2 = 6, a_3 = -4, a_4= 1

- Remove the trend i.e.,

$$y_1[n] = y[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y[n-k] \tag{3}$$

$$\widehat{y}[n] = y_1[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y_1[n-k] \tag{4}$$

where $2N+1$ corresponds to the average pitch period over a longer segment of speech

- The trend removed signal $\widehat{y}(n)$ is termed as ZFFS.

The normalized autocorrelation peak strength (NAPS) of ZFFS [13, 17] is computed and this is found to be high for the foreground speech regions and low for the background noise regions. The NAPS was used as a feature for foreground speech segmentation in [13] and the same feature is being used in this work. The NAPS of ZFFS can be obtained by first computing the autocorrelation of a 25 ms frame of ZFFS along with a frame shift of 10 ms. The value of the first largest peak of the autocorrelation sequence (excluding the central peak) is considered by normalizing with respect to the central peak. The ZFFS and the NAPS for a segment of outdoor speech is shown in Figure 1 (b) and (c), respectively.

It can be observed that the NAPS of ZFFS shows a high value for the foreground regions and relatively lower values for the background region. However, it is difficult to set the threshold value to segment the foreground and background regions. Alternatively, the segmentation can be performed by using a non-linear mapping function. This non-linear mapping is applied to further exaggerate the discrimination between the two regions. The non-linear mapping function basically consists of an exponential function and is defined by the following equation.

$$P_m = \frac{1}{1+e^{-(P_s-\theta)/\tau}} + \alpha \tag{5}$$

where, P_m is non-linearly mapped value, P_s is the NAPS of ZFFS, θ, τ are the slope parameters and α is the offset that defines the minimum value of the function. θ is the main threshold and its effect on the task is explained in the Section 3. The τ is set to a very low value to obtain almost a zero or one mapping and α is set to zero. $P_m = 1$, if $P_s > \theta$, otherwise $P_m = 0$. The non-linear mapped value of NAPS of ZFFS is shown in Figure 1(d). It can be seen that the foreground and background segments are separated, where the foreground regions are indicated by the label one while the background regions are indicated by the label zero.

The foreground speech segmentation is then applied on both the indoor and outdoor speech segments and can be seen in Figure 2. It can be observed that the NAPS of ZFFS has different value for the indoor and outdoor segments mainly in the background regions marked as continuous black line in the figure. This difference in the value of the NAPS of ZFFS gives an idea that if the background regions of the indoor and outdoor speech are processed and appropriate features are extracted from them, the classification of indoor and outdoor speech can be performed. The difference in the value of NAPS is due to the fact that the outdoor segments contain some noise in the background whereas the indoor segments do not have the kind of noises present in the outdoor speech. In most cases the signal energy in the background of the indoor segments is very low compared to the outdoor segments.

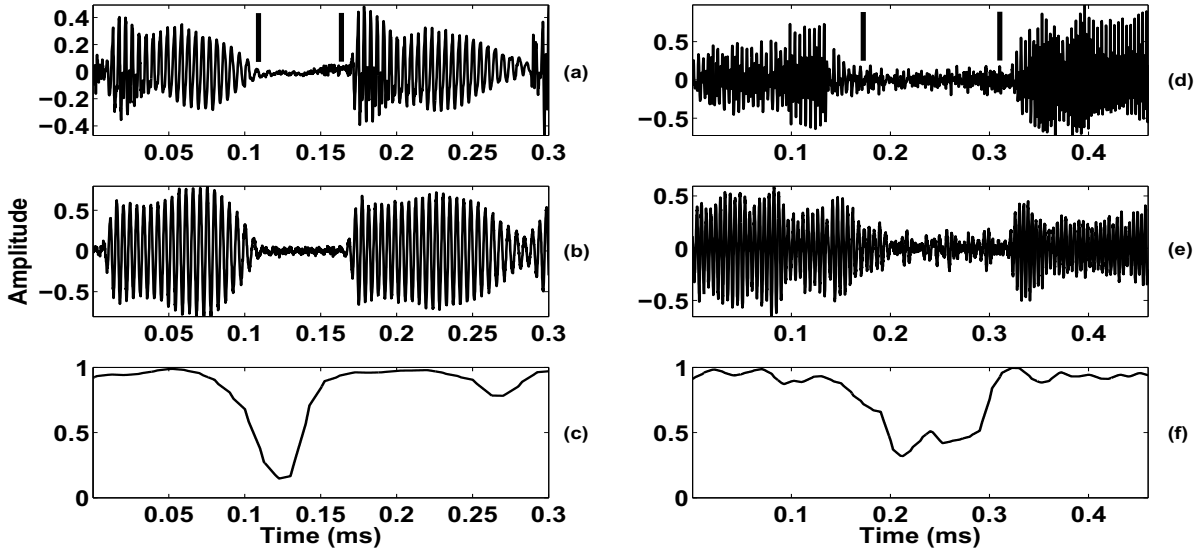

Figure 2: *Illustration of the difference of the background regions for indoor and outdoor segments (a) indoor speech (b) ZFFS of indoor speech (c) NAPS of ZFFS of indoor speech (d) outdoor speech (e) ZFFS of outdoor speech (f) NAPS of ZFFS of outdoor speech. A background region is marked with the continuous black line*

2.1. Classification using Features from the Background Regions

The background regions are considered for the classification purpose. The Mel Frequency Cepstral Coefficients (MFCCs) features are extracted from such regions. In order compute the MFCCs a frame size of 25 ms along with a frame shift of 10 ms is used. Initially the manually marked background regions are considered for feature extraction. The obtained features are used for training the Support Vector Machine (SVM) classifier [15, 16] with one class belonging to indoor and the other class belonging to outdoor. The models created using the MFCC features for indoor and outdoor are then subsequently used for classification of a given segment. However, during the testing phase the audio is divided into 5 s non-overlapping segments. Each segment is further subjected through foreground and background segmentation. The MFCC features are extracted from background regions for every frame. The MFCC features extracted are then tested with the trained models. If the number of MFCC features classified into a particular class exceed a threshold then that segment of 5 s is classified into that particular class.

3. Results and Discussion

The experiments were performed on the broadcast audio collected from Indian news channels (english) at a sampling rate of 8 kHz. Totally 900 audio segments are considered for training and testing purpose, where each audio segment is having the length of 5 s. Out of that, 100 segments are used for training, 100 segments are used for validation while the remaining 700 samples are used for testing. The training is done using a cross validated framework, where initially the data is divided into training set, validation set and testing set. A 4-fold cross validation was used for setting the optimal parameters of the classifier using the validation set. The parameters are varied according to a grid search. All the experiments using SVM were carried out using the libSVM [18] with a radial-basis function (RBF) kernel of the form,

$$K(x, y) = exp(-Y||x - y||^2) \quad (6)$$

The width parameter Y and the cost parameter c are the parameters which are optimized. The testing is performed based on the optimized parameters. The threshold θ for the non-linear mapping function (equation 5) is a tunable parameter and the results are presented by varying this parameter. The Figure 3 shows the results using the SVM classifier along with the variation of the parameter θ of the non-linear mapping function.

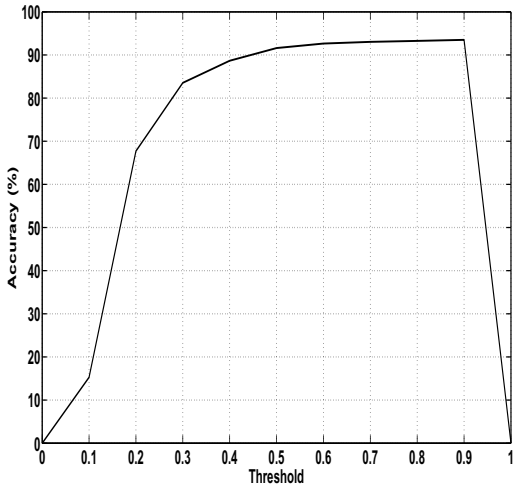

Figure 3: *Results in terms of the overall accuracy of indoor/outdoor audio classification with variation of the threshold θ of the non-linear mapping function*

Table 1: **Confusion Matrix for indoor/outdoor classification.**

Confusion Matrix	Indoor	Outdoor
Indoor	700	0
Outdoor	91	609

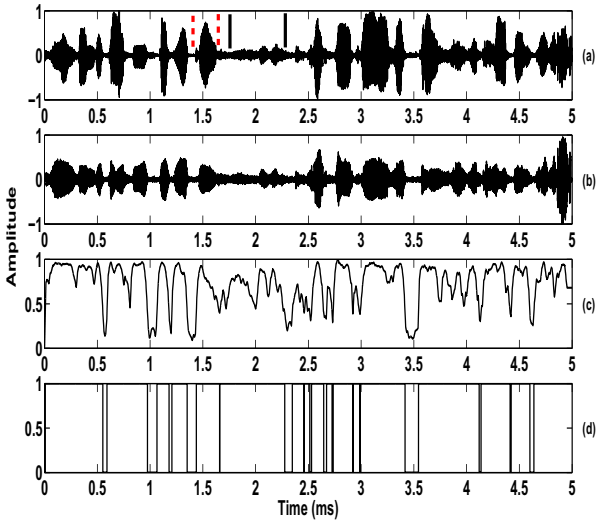

Figure 4: *Illustration of foreground speech segmentation on a sample of 5 s of outdoor speech which contains unwanted sources in the background as speech. (a) Speech signal (b) ZFFS of speech (c) Normalized autocorrelation peak strength (NAPS) of ZFFS (d) Non-linearly mapped value of (c) in which the foreground regions are mapped to one and the background regions are mapped to zero. A sample of a foreground region is marked with the dotted red line and a background region is marked with the continuous black line*

The results in the Figure 3 reveal that on increasing the threshold θ, the performance improves and achieves a maximum of 93.27% for a θ value of 0.9. The reason for this is because when the threshold is set high, the segmented background regions increase for a particular audio file. This in turn gives more number of features for testing on the classifiers which gives the classifier more variability to decide whether the particular audio file is an indoor or an outdoor sample. If lesser number of regions are segmented as background, the features obtained from those background regions for a particular audio file may not give enough information about the particular background present in the audio file which leads to errors in the classification task as seen in the figure with the decrease in the threshold value.

A confusion matrix of the classification is also presented in table 1. The table indicates that the indoor speech samples are classified almost 100% correctly. The decrease in the overall performance is due to the misclassification of the outdoor speech samples as indoor. The reason for this is because some of the outdoor audio files contain speech in the background. An illustration of such a file can be seen in Figure 4. A sample foreground region is in between the dotted red lines and a sample background region is in between the continuous black line. It can be noted that for this particular audio file, speech is present in the background as opposed to other types of background noises. The presence of the speech in the background causes some of the background regions to be segmented as foreground. This causes the remaining background which does not contain speech but silence to be used for the classification task since these segments will be classified as background only. These silence background regions will have a nature as the background regions of the indoor audio and hence the corresponding outdoor audio file will be classified as an indoor audio file.

4. Conclusion and Future Work

This work explored the use of foreground speech segmentation for indoor/outdoor classification. The NAPS of ZFFS of speech was used as a feature for the foreground speech segmentation. The background regions obtained are considered for the classification task wherein the MFCC features are extracted from them. The classification is performed using SVM classifier. The results show that the foreground speech segmentation task does aid in the indoor versus outdoor classification and this task can be used further as a preprocessing task for applications such as speech recognition.

It was observed that some of the background regions contained background speech which results in error for the foreground speech segmentation module and this in turn affects the performances of the indoor versus outdoor classification. The future work would involve determining the type of background noise present in outdoor speech and appropriate features can be used for the segmentation process in order to improve the overall task.

5. Acknowledgements

This work is part of the project titled *ARTICULATE +: A system for automated assessment and rehabilitation of persons with articulation disorders* funded by the Ministry of Human Resource Development, Govt. of India under IMPRINT. The authors would also like to thank Udeshna Deka for her assistance in the data preparation.

6. References

[1] J. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," *Communications Of the ACM*, vol. 43, no. 2, pp. 64–70, February 2000.

[2] P. Woodland, "The development of the HTK broadcast news transcription system: An overview," *Speech Communication*, vol. 37, no. 1-2, pp. 47–67, May 2002.

[3] L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, vol. 38, no. 1-2, p. 213230, September 2002.

[4] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, no. 1, pp. 5–20, 2000.

[5] S. Wegmann, P. Zhan, and L.Gillick, "Progress in broadcast news transcription at Dragon Systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1999, pp. 33–36.

[6] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[7] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech communication*, vol. 37, no. 1, pp. 89–108, 2002.

[8] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus, "Large vocabulary continuous speech recognition of Broadcast News–The Philips/RWTH approach," *Speech Communication*, vol. 37, no. 1, pp. 109–131, 2002.

[9] M. J. Gales, P. Woodland, H. Y. Chan, D. Mrva, R. Sinha, S. E. Tranter *et al.*, "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1513–1525, 2006.

[10] D. A. Van Leeuwen, J. M. Kessens, E. Sanders, and H. Van Den Heuvel, "Results of the n-best 2008 dutch speech recognition evaluation." in *INTERSPEECH*, 2009, pp. 2571–2574.

[11] H. Kamper, F. De Wet, T. Hain, and T. Niesler, "Capitalising on north american speech resources for the development of a south african english large vocabulary speech recognition system," *Computer Speech & Language*, vol. 28, no. 6, pp. 1255–1268, 2014.

[12] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[13] K. Deepak, B. D. Sarma, and S. R. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[14] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1204–1218, 2016.

[15] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1602–1613, Nov 2008.

[16] K. S. Srinivas and K. Prahallad, "An fir implementation of zero frequency filtering of speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2613–2617, 2012.

[17] B. K. Khonglah and S. R. M. Prasanna, "Speech/music classification using speech-specific features," *Digital Signal Processing*, vol. 48, pp. 71–83, 2016.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.