IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

A Recurrent Neural Network Approach to Audio Segmentation for Broadcast Domain Data

Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{pablogj, ivinalsb, ortega, amiguel, lleida}@unizar.es

Abstract

This paper presents a new approach for automatic audio segmentation based on Recurrent Neural Networks. Our system takes advantage of the capability of Bidirectional Long Short Term Memory Networks (BLSTM) for modeling temporal dynamics of the input signals. The DNN is complemented by a resegmentation module, gaining long-term stability by means of the tied-state concept in Hidden Markov Models. Furthermore, feature exploration has been performed to best represent the information in the input data. The acoustic features that have been included are spectral log-filter-bank energies and musical features such as chroma. This new approach has been evaluated with the Albayzín 2010 audio segmentation evaluation dataset. The evaluation requires to differentiate five audio conditions: music, speech, speech with music, speech with noise and others. Competitive results were obtained, achieving a relative improvement of 15.75% compared to the best results found in the literature for this database.

Index Terms: audio segmentation, recurrent neural networks, LSTM, broadcast data

1. Introduction

Due to the increase of multimedia content and the generation of large audiovisual repositories, the need for automatic systems that can analyze, index and retrieve information in a fast and accurate way is becoming more and more important. Given an audio signal, the goal of audio segmentation is to obtain a set of labels in order to separate that signal into homogeneous regions and classify them into a predefined set of classes, e.g., speech, music or noise. This task is also important in other applications of speech technologies, such as automatic speech recognition (ASR) or speaker diarization, where an accurate labeling of audio signals can improve the performance of these systems in real-world environments.

Audio segmentation systems can be divided into two main groups depending on how the segmentation is performed: segmentation-and-classification systems and segmentation-by-classification systems.

- **Segmentation-and-classification**: these systems perform the segmentation task in two steps. First, boundaries that separate segments belonging to different classes are detected using a distance metric. Then the system classifies each delimited segment in a second step. Several distance metrics have been proposed in the literature: Bayesian Information Criterion (BIC) [1] [2], Generalized Likelihood Ratio (GLR) [3], or the Kullback Leibler (KL) distance [4] are some examples.
- **Segmentation-by-classification**: in this group of systems the segmentation is produced directly as a sequence of decisions over the input signal. A set of well-known machine learning classification techniques have been used in this task with good results, such as Gaussian Mixture Models (GMM) [5], Neural Networks (NN) [6] [7], Support Vector Machines (SVM) [8], or decision trees [9]. A factor analysis approach is proposed in [10], where a compensation matrix is computed for each class.

In both approaches to audio segmentation it is usual that the original segmentation boundaries are refined by a resegmentation model. In this way, sudden changes in the labels can be prevented. Some resegmentation strategies rely on hidden Markov models [11] or smoothing filters [12].

The segmentation task is specially challenging when dealing with broadcast domain content because such documents contain different audio sequences with a very heterogeneous style. Different speech conditions and domains can be found, from telephonic quality to studio recordings to outdoors speech with different noises overlapped. Background music and a variety of acoustic noise effects are likely to appear as well. In this context, the technological evaluations Albayzín incorporated a segmentation task for broadcast news environments in 2010 [13]. This is the task we are focusing on in this paper.

The remainder of the paper is organized as follows: a theoretical background on LSTM networks and its applications is introduced in section 2. Section 3 describes our novel RNN-based segmentation system. Section 4 briefly introduces the experimental setup, describing the Albayzín 2010 evaluation dataset and presents all the experimental results obtained with our system. Finally, a summary and the conclusions are presented in section 5.

2. LSTM networks

Neural Networks are a powerful modeling tool for non linear dependencies that, since the early 2010s, have been increasingly applied to speech technologies [14] [15] [16]. One of the main disadvantages of traditional feed-forward neural networks when dealing with temporal series of information is that they process every example independently. However, Recurrent Neural Networks (RNNs) are able to capture temporal dependencies introducing feedback loops between the input and the output of the neural network. The long short term memory (LSTM) [17] networks are a special kind of RNN with the concept of memory cell. This cell is able to learn, retain and forget information [18] in long dependencies.

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the 2015 FPI fellowship, the project TIN2017-85854-C4-1-R and Gobierno de Aragón /FEDER (research group T36_17R)

10.21437/IberSPEECH.2018-19

This capability becomes very useful to carry out long and short term analysis simultaneously. LSTM networks have been modified combining two of them in a Bidirectional LSTM (BLSTM) network. One processes the sequence in the forward direction while the other one processes the sequence backwards. This way the network is able to model causal and anticausal dependencies for the same sequence.

LSTM ans BLSTM networks have been successfully applied to sequence modeling tasks in speech technologies such as ASR [19] [20], language modeling [21], speaker verification in text-dependent systems [22] or machine translation [23].

3. System description

Our proposed system is based on the use of RNNs to classify each frame in the audio signal. We have opted for a segmentation-by-classification solution, combining the RNN with a resegmentation module to get smoothed segmentation hypotheses. The three different blocks of our segmentation system (feature extraction, an RNN based classifier, and the final resegmentation module) are described below.

3.1. Feature extraction

The main features for the neural network consist of log Mel filter bank energies and the log-energy of each frame. Additionally, we combine Mel features with chroma features [24], due to its capability to capture melodic and harmonic information in music while being robust to changes in tone or instrumentation. All these features are computed every 10 ms using a 25 ms window. In order to take into account the dynamic information of the audio signal, first and second order derivatives of the features are computed. Finally, feature mean & variance normalization is applied.

3.2. Recurrent Neural Network

This is the core of our segmentation system. Its task is to classify each audio frame as belonging to one of the predefined acoustic classes. The neural architecture proposed can be observed in Fig. 1. As shown, it is composed by two stacked BLSTM layers with 256 neurons each. The outputs of the last BLSTM layer are then independently classified by a linear perceptron, which shares its values (weights and bias) for all time steps. Training and evaluation are performed with limited length sequences (3 seconds, 300 frames), limiting the delay of dependencies to take into account. However, the neural network emits a segmentation label per every frame processed at the input (every 10 ms, in our case).

The neural network has been trained using exclusively the train subset of the Albayzín 2010 database [13], which consists of 58 hours of audio sampled at 16 KHz. However, 15% of the train subset will be reserved for validation, which makes a total of 49 hours of audio for training and 9 hours for validation. Adaptative Moment Estimation (Adam) optimizer is chosen due to its fast convergence properties [25]. Data will be shuffled in each training iteration seeking to improve model generalization capabilities.

Despite our system consisting of an additional resegmentation module, the RNN itself is able to emit a segmentation hypothesis. This way, we can say that the first two blocks of our system can fully work as a segmentation system. All the neural architectures in this paper have been evaluated using the PyTorch toolkit [26].

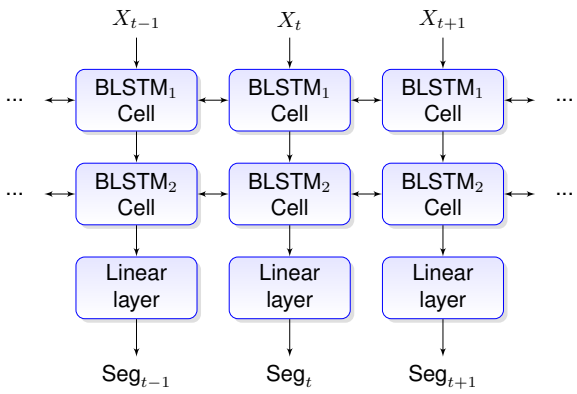

Figure 1: *BLSTM based neural architecture description. X_i represents the input features for the frame i. Seg_i is the segmentation label for frame i*

3.3. Resegmentation module

RNNs have high capabilities to model temporal dependencies but its output may contain high frequency transitions which are unlikely to occur in signals which have a high temporal correlation such as human speech or music. With the objective of avoiding sudden changes in the segmentation process, we incorporate a resegmentation module to our system. In our case, it is implemented as a hidden Markov model where each acoustic class is modeled through a state in the Markov chain. Every state is represented by a multivariate Gaussian distribution with full covariance matrix. This block gets as input the pseudo log-likelihood for each class from the neural network and its corresponding segmentation hypotheses. No more a priori information is required because statistical distributions are estimated using the hypothesized labels for each file.

Input information is given every 10 ms, which can result in a noisy estimation of class boundaries. In order to reduce temporal resolution, scores are down-sampled by a factor L using an L order averaging zero-phase FIR filter to avoid distorting phase components [27]. First and second order derivatives of the scores are taken into account when computing the resegmentation. Additionally, each of the states in the Markov chain will consist of a left-to-right topology of a number N_{ts} of tied states sharing the same statistical distribution. This way, by modifying N_{ts} we can force the minimum length of a segment before a change in the acoustic class happens. Taking all this information into account, this minimum segment length forced by our resegmentation module can be computed as follows:

$$T_{min} = T_s L N_{ts} \tag{1}$$

where, T_s is the sampling period of the neural network output (10 ms in our case), L is the down-sampling factor and N_{ts} is the number of tied states used in the Markov model.

4. Experimental setup and results

4.1. Database and metric description

The database consists of broadcast news audio in Catalan. The full database includes 87 hours of audio sampled at 16 KHz and divided in 24 files. The database was split into two parts: two thirds of the total amount of data are reserved for training, while the remaining third is used for testing. Five acoustic classes were defined for the evaluation. The classes are distributed as follows: 37% for clean speech (sp), 5% for music (mu), 15% for speech over music (sm), 40% for speech over noise (sn) and 3% for others (ot). The class "others" is not evaluated in the final test. A more detailed description of the Albayzín 2010 audio segmentation evaluation can be found in [13].

The main metric we will be using for evaluating our results is the the Segmentation Error Rate (SER), inspired by the NIST metric for speaker diarization [28]. This metric can be interpreted as the ratio between the total length of the incorrectly labeled audio and the total length of the audio in the reference. Given the dataset to evaluate Ω, each document is divided into continuous segments and the segmentation error time for each segment n is defined as:

$$\Xi(n) = T(n)[\max(N_{ref(n)}, N_{sys}(n)) - N_{correct}(n)] \quad (2)$$

where $T(n)$ is the duration of the segment n, $N_{ref}(n)$ is the number of reference classes that are present in segment n, $N_{sys}(n)$ is the number of system classes that are present in segment n and $N_{correct}(n)$ is the number of reference classes that are present in segment n and were correctly assigned by the segmentation system. This way, the SER is computed as follows:

$$\text{SER} = \frac{\sum_{n\in\Omega}\Xi(n)}{\sum_{n\in\Omega}(T(n)N_{ref}(n))} \quad (3)$$

Alternatively, the metric originally proposed for the Albazín 2010 evaluation will also be taken into account. This metric represents the relative error averaged over all the acoustic classes:

$$\text{Error} = \text{mean}_i \frac{\text{dur}(\text{miss}_i) + \text{dur}(\text{fa}_i)}{\text{dur}(\text{ref}_i)} \quad (4)$$

where $\text{dur}(\text{miss}_i)$ is the total duration of all miss errors for the ith acoustic class, $\text{dur}(\text{fa}_i)$ is the total duration of all false alarm errors for the ith acoustic class, and $\text{dur}(\text{ref}_i)$ is the total duration of the ith acoustic class according to the reference. A collar of $\pm$1s around each reference boundary is not scored in both cases, SER and average class error, to avoid uncertainty about when an acoustic class begins or ends, and to take into account inconsistent human annotations.

4.2. Experimental results

For the experimental evaluation of our system, different front-end configurations were assessed. The starting point of our feature space exploration consists of a simple 32 log Mel filter bank. Our next step was increasing the frequency resolution by using a higher number of analysis bands in the filter bank, testing 64, 80 and 96 bands. Chroma features were incorporated in order to help our system to discriminate classes that contain music. Eventually, first and second order derivatives were computed to take into account dynamic information in the audio signal.

Table 1: *SER, error per class and average error for BLSTM segmentation-by-classification system on the test partition for different feature configurations (Mel: log Mel filter bank, Chr: chroma, $\Delta + \Delta\Delta$: 1st and 2nd order derivatives)*

Feats	SER	Class Error(%)				Avg
		mu	sp	sm	sn	
32 Mel	17.67	18.09	30.89	33.07	35.14	29.30
64 Mel	17.47	17.98	31.45	31.38	34.60	28.85
80 Mel	**16.87**	**18.14**	**29.63**	**30.23**	**33.45**	**27.86**
96 Mel	17.33	18.07	30.81	31.48	33.94	28.58
80 Mel + Chr	16.14	17.12	30.13	26.57	31.66	26.37
80 Mel + Chr + $\Delta+\Delta\Delta$	**15.91**	**16.28**	**28.82**	**26.32**	**31.94**	**25.84**

Results obtained on the test partition for the different front-end configurations using our BLSTM segmentation-by-classification system are presented in Table 1 in terms of SER and the Albayzín evaluation metric. When the number of analysis bands is increased, we can appreciate that the SER decreases, reaching its minimum using 80 bands. However, it can also be seen that using a higher number of bands can affect the system performance. This is the case of the 96 bands configuration, that increases its error compared to the 80 bands configuration. We can notice that, by incorporating chroma features, the error in the class "Speech over music" decreases significantly when compared to the 80 Mel coefficient configuration, with a relative improvement of 12.10%. This is due to the capabilities of chroma features to capture musical dependencies, which helps our system discriminate this class in a more accurate way. The best result for this set of experiments is obtained using the first and second order derivatives of the log Mel filter bank and the chroma features, achieving a SER of 15.91%, which is equivalent to an average class error of 25.84%. The performance of our segmentation system using the resegmentation module is evaluated in the following set of experiments.

Aiming to illustrate how the system performance is influenced by the inertia imposed by the resegmentation module, Fig. 2 shows the scatter plot of the relative improvement in performance versus the minimum segment length (T_{min}) for different values of the down-sampling factor L. It can be seen that configurations that perform better have a minimum segment length between 0.5 and 1.5 seconds, which is in the order of magnitude of the 2 seconds collar applied in the evaluation.

The results on the test partition of the full segmentation system combining the BLTSMs and the resegmentation module for the best front-end configuration evaluated (80 Mel + chroma + derivatives) and for different values of the down-sampling factor, L, and the minimum segment length, T_{min}, are shown in Table 2 in terms of SER and the Albayzín evaluation metric. If we compare the best result in this Table with the best result in Table 1, it can be seen that, by incorporating

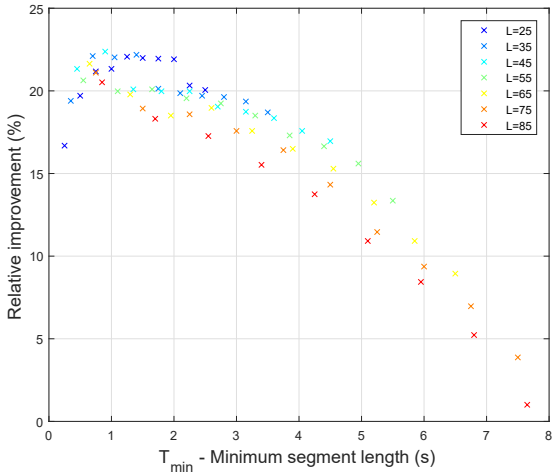

Figure 2: *Relative improvement using the resegmentation module for the best feature configuration versus minimum segment length forced by the system*

Table 2: *SER, error per class and average error for BLSTM-HMM segmentation-by-classification system over test partition for the best feature configuration and different values of the down-sampling factor, L, and minimum segment length, T_{min}*

L, T_{min}	SER	Class Error(%)				Avg
		mu	sp	sm	sn	
25, 1.25s	12.49	14.55	21.99	19.08	24.88	20.13
35, 1.4s	12.48	14.31	22.26	18.70	25.10	20.10
45, 0.9s	**12.46**	**14.19**	**22.14**	**18.82**	**25.04**	**20.05**
55, 0.55s	12.57	16.12	22.00	18.94	24.95	20.50

the resegmentation module, the error is reduced significantly, getting a relative improvement of 21.68% in terms of SER. It can also be observed that, as long as the T_{min} value used in the configuration stays in the order of magnitude of 1 second, the performance of the system is not highly affected by the variations in the down-sampling factor. The SER metric for the four parameter configurations evaluated goes from 12.46% to a 12.57%, which makes an absolute difference of only 0.11% between the best and the worst case. This way, by forcing a certain amount of inertia in the output of the neural network, the system is able to achieve a SER of 12.46%, decreasing significantly the error when compared to the output of the RNN.

Finally, Table 3 shows the results obtained in the Albayzín 2010 database by different systems already presented in the literature. The winner team of the Albayzín 2010 evaluation proposed a segmentation-by-classification approach based on a hierarchical GMM/HMM including MFCCs, chroma and spectral entropy as input features [29]. The best result in this database so far uses a solution based on Factor Analysis combined with a Gaussian back-end and MFCCs with 1st and 2nd order derivatives as input features [10]. When compared with this result, it can be seen that our BLSTM-HMM system performs better in all the acoustic classes. This error reduction is equivalent to a relative improvement of 15.24% in terms of SER and a 15.75% in terms of the average class error. The difference in the class "speech over music" is specially significant (23.60% vs 18.82%) with a relative improvement of 20.25%.

Table 3: *Results obtained on the Albayzín 2010 test partition for different systems proposed in the literature compared to our BLSTM-HMM system*

System	SER	Class Error(%)				Avg
		mu	sp	sm	sn	
Eval winner [29]	19.30	19.20	39.50	25.00	37.20	30.30
FA HMM [10]	14.70	18.80	23.70	23.60	29.10	23.80
BLSTM HMM	**12.46**	**14.19**	**22.14**	**18.82**	**25.04**	**20.05**

5. Conclusions

A new approach for audio segmentation based on RNNs is presented in this paper proving the capabilities of this kind of models in the audio segmentation task, achieving the best result so far in the Albayzín 2010 database. A segmentation-by-classification scheme has been followed, combining a classification system, which is mainly made of 2 BLSTM layers, with an smoothing back-end implemented through a Hidden Markov Model. Several front-end configurations were evaluated, proving the capabilities of chroma features for capturing musical structures when compared to a perceptual Mel filter bank. The combination of BLSTM and HMM has been proven to be appropriate, reducing significantly the system error by forcing a minimum segment length for the segmentation labels. Competitive results have been obtained with this new approach, resulting in a relative improvement of 15.75% when compared to the best result in the literature so far.

Regarding our contributions, front-end configuration seems to have a big impact in this task, specially when classifying classes that contain music. Just by modifying the input features we have achieved a significant improvement in the performance of our system. Furthermore, the introduction of RNNs in the audio segmentation task has been proven to be successful, improving the results obtained so far with traditional statistical models such as GMM/HMM or factor analysis. In future work we intend to improve even more these results by introducing more complex neural architectures after the BLSTM layers.

6. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

[1] S. Chen, P. Gopalakrishnan *et al.*, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8, 1998, pp. 127–132.

[2] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 266–276, 2006.

[3] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech communication*, vol. 32, no. 1-2, pp. 111–126, 2000.

[4] 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*, vol. 1997, 1997.

[5] A. Misra, "Speech/nonspeech segmentation in web videos," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] H. Meinedo and J. Neto, "A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[7] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.

[8] G. Richard, M. Ramona, and S. Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2, 2007, pp. II–461.

[9] Y. Patsis and W. Verhelst, "A speech/music/silence/garbage/classifier for searching and indexing broadcast news material," in *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, 2008, pp. 585–589.

[10] D. Castán, A. Ortega, A. Miguel, and E. Lleida, "Audio segmentation-by-classification approach based on factor analysis in broadcast news domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 34, 2014.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech communication*, vol. 40, no. 3, pp. 351–363, 2003.

[12] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 203–211.

[13] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 1, 2011.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[15] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[16] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[19] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[21] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[22] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[24] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 15–18.

[25] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[26] A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, L. Antiga, and Z. Devito, "Automatic differentiation in PyTorch," *Advances in Neural Information Processing Systems 30*, pp. 1–4, 2017.

[27] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 988–992, 1996.

[28] NIST, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," (Melbourne 28-29 May 2009).

[29] A. Gallardo Antolín and R. San Segundo Hernández, "UPM-UC3M system for music and speech segmentation," 2010.