Investigating the Use of Semi-Supervised Convolutional Neural Network Models for Speech/Music Classification and Segmentation

David Doukhan and Jean Carrive
French National Institute of Audiovisual (Ina)
Paris, France
Email: ddoukhan@ina.fr, jcarrive@ina.fr

Abstract—**A convolutional neural network architecture, trained with a semi-supervised strategy, is proposed for speech/music classification (SMC) and segmentation (SMS). It is compared to baseline machine learning algorithms on three SMC corpora and demonstrates superior performances, associated to perfect media-level speech recall scores. Evaluation corpora include speech-over-music segments with durations varying between 3 and 30 seconds. Early SMS results are presented. Segmentation errors are associated to musical genres not covered in the training database, and/or with close to speech acoustic properties. These experiments are aimed to help the design of novel speech/music annotated resources and evaluation protocols, suited to TV and radio stream indexation.**

Keywords–Speech/music discrimination; Audio segmentation; Convolutional Neural Networks; Music Information Retrieval; Multimedia Indexation.

I. Introduction

Speech/Music classification (SMC) task consists in predicting if a given audio excerpt contains speech or music. The excerpts are supposed to be pure and contain either speech or music. Classified excerpts may have variable durations. Longer excerpt durations are known to make classification tasks easier. Speech excerpts are generally defined as containing *spoken speech*: this includes speech alone, superposed voices, and speech over music. Music excerpts are defined as containing instruments, instruments mixed with lyrics, or *a cappella* vocals. These technologies attracted much interest for the management of large multimedia collections: selection of optimal audio compression strategy at Swedish radio [1], as well as tag correction in Deezer's catalog [2].

Speech/Music segmentation (SMS) task consists in splitting audio streams into pure speech and pure music segments [3]. SMS algorithms are often based on frame-level SMC procedures, followed by a post-processing step (mean filtering, dynamic programming, etc.). SMS is a pre-processing stage required for several higher level indexation tasks such as speech and speaker recognition, song and musical genre recognition. Consequently, their development has received considerable attention from speech analysis and music information retrieval communities, illustrated by several evaluation campaigns, e.g. ESTER [4], Albayzín-2014 [5] or MIREX 2015 [6].

This paper presents the ongoing research on SMC and SMS tasks carried out at French National Institute of Audiovisual (Ina). Ina is a public institution in charge of the preservation, digitization, distribution and dissemination of the French audiovisual heritage. Ina's archives represent 70 years of radio and 60 years of TV programs, for a total of 15 million hours. The integration of SMS technologies in Ina workflows would allow a fast localization of interest areas within audio recordings, and address several identified needs. SMS may help speeding up media descriptions processes, which are performed manually by professional archivists. Manual media description is expensive, and associated to variable levels of detail: TV broadcast news are described with greater details than early radio collections. Consequently, SMS may ease the browsing and exploitation of under-documented archive contents. Latest identified use-case is music track segmentation, aimed at detecting and measuring the duration of musical tracks, in order to calculate the amount of royalties to be paid to rights collection societies.

The work presented in this paper is a preliminary study on SMC and SMS issues, using Convolutional Neural Networks (CNN). It is motivated by the excellent results reported with these architectures on MIREX 2015 SMC and SMS tasks, consisting in classifying 30-seconds long pure music and pure speech audio files. MIREX 2015's best SMC results were obtained using CNNs trained using fully supervised procedures: a MFC-based model with 1 convolutionnal layer [7], and a CQT-based model with 3 convolutionnal layers [8]. The main contribution of this paper is the description of MFC-based CNN's performances on publicly available datasets, using shorter audio segments (from 3 to 30 seconds), as well as speech-over-music excepts. Another major contribution is the proposal of an unsupervised SMC training strategy used in the first layer of the network, allowing to obtain visually relevant audio classification features.

This paper is structured as follows. Section II describes the audio feature extraction process. Section III describes the proposed CNN architecture and training strategy. Section IV and V describe the corpora used for SMC evaluation, and the corresponding results. Section VI describes the early results obtained on the SMS task. Section VII provides a summary of the results obtained, and introduces our future work.

II. Feature extraction and normalization

Audio excerpts are downsampled to 16KHz mono signals. Mel-frequency cepstrums (MFC) corresponding to 40 Mel bands are extracted from 20 ms frames, sampled with a 10 ms time step. Adjacent frames are concatenated using a contextual length varying between 1 (no context, 40 dimensions) and 50 (context of 500 ms, 2000 dimensions).

The resulting features are firstly normalized at the media level through a cepstral mean subtraction and a standard deviation division process. Similarly to *patch normalization* procedures, mean and standard deviation are also computed for each feature vector, and used to perform local mean subtraction and standard deviation division process.

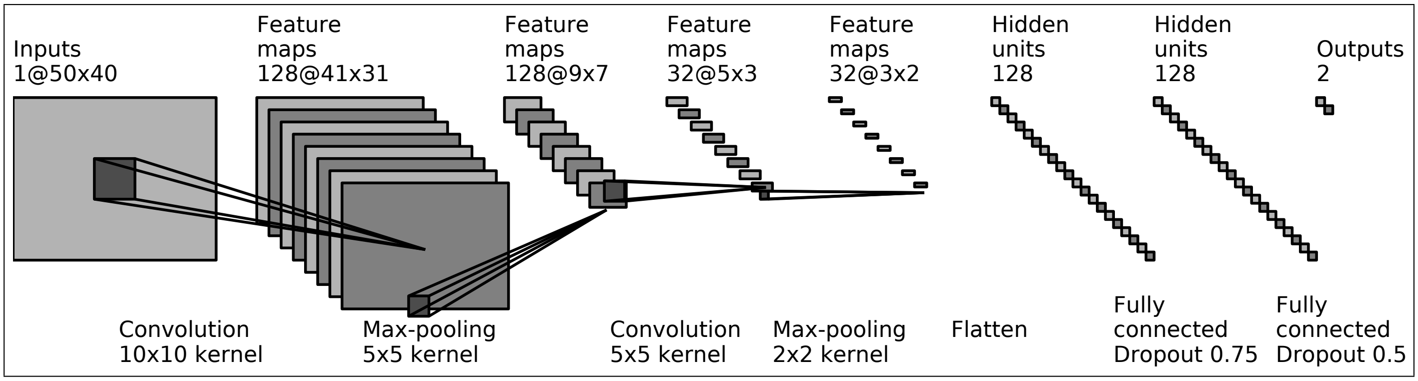

Figure 1. Proposed convolutional neural network architecture for frame-level speech/music classification

III. Semi-Supervised Convolutional Neural Network Model

CNNs are a type of feed-forward artificial neural network, in which the response of neurons to stimuli is triggered by convolution operations [9]. The weights are organized into a set of *filters* allowing the detection of localized spatio-temporal patterns, behaving like task-oriented feature extractors. Pattern shift and space invariance is achieved through the use of *pooling* operations.

In the scope of this study, several CNN architectures and training strategies were implemented using Tensorflow [10] and compared using variable number of convolutional and densely connected layer, filter shapes, pooling strategies, regularization methods. Figure 1 shows the structure of the CNN model associated to the best performances, reported in the next sections.The model input is composed of 50 MFC frames, corresponding to a 500 milliseconds temporal resolution. The frames are first processed by two convolutional layers (128 10×10 and 32 5×5 filters) reduced by max-pooling layers (5×5 and 2×2). These layers are followed by two densely connected layers of 128 neurons, which are associated to dropout rates of 0.75 and 0.5 [11]. Rectified linear unit (ReLU) activation functions are used between layers. The output layer is composed of two neurons, normalized using a softmax function, corresponding to the detection of music or speech. The first convolutional layer of the network is trained using an unsupervised procedure based on Spherical K-Means and ZCA whitening [12], with K being the number of filters of the first layer. Filters obtained through this procedure are associated to unit l^2 norm, which avoids irregular neural response magnitude, and prevents overfitting. First layer's filters are associated to visually relevant features (vertical, horizontal and diagonal patterns) illustrated by figure 2. Visual relevance of these shapes is also a desirable property (see [2] for filter shapes associated to fully supervised regularization strategies). The first layer stays constant during the training procedure, and remaining layers are trained using supervised Adaptive Moment Estimation (Adam) gradient descent optimization algorithm [13], for a maximal amount of 30 *epochs*, and are stopped when the accuracy on the training set reaches 99.9%.

IV. Speech/Music classification corpora

Three publicly available SMC corpora were used, each of them being associated to specific speaking styles, musical genres, and track durations.

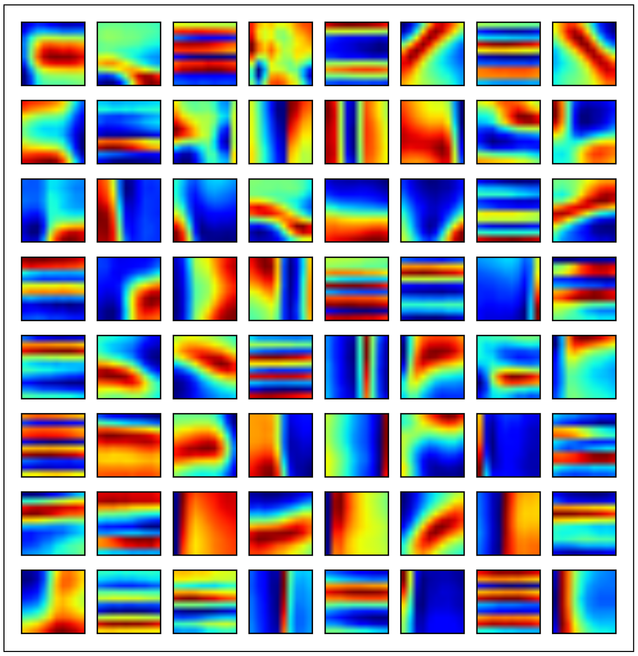

Figure 2. 10×10 Filters used in the first layer of the CNN

GTZAN music/speech collection [14] contains 120 tracks, each 30 seconds long. It contains 60 examples of speech and 60 examples of music. Speech samples cover various languages, accents and contexts: comic, radio, films, interviews, dialogues, advertisements, news, story-telling, etc. Music samples include several genres, including vocals without instruments.

Scheirer-Slaney Music/Speech corpus [15] contains 240 15-seconds tracks. These excerpts were collected from radio streams, which is a context representative of Ina's use case. It is composed of 140 examples of speech (English and Spanish) and 100 examples of music (with and without vocals). An interesting feature of this corpus is that 60 speech examples correspond to speech-over-music, which is supposed to increase the difficulty of classification and segmentation tasks.

Music, Speech, and Noise Corpus (MUSAN) [16] contains about 60 hours of speech and 42 hours of music. 20 hours of the speech material are full audiobook chapters, read in 12 languages, and correspond to clean read speech. The 40 remaining hours of speech material are composed of US government hearings, committees and debates, containing background noise (noisy spontaneous and prepared speech). Music material is provided with annotations related to performers and musical genres. 3 seconds long segments were extracted from MUSAN corpus through a procedure aimed at measuring the ability of the models to discriminate speech versus music on small segments and providing controlled variations on the training and testing examples. MUSAN's speech material is composed of 2 randomly selected segments per track in US government material, and 3 segments per audiobook, amounting to 1024 segments, for a duration of 9 minutes for each speech category. Music material is composed of some 6 randomly extracted segments per artist, amounting to 924 segments obtained from 252 different artists. Automatic energy-based procedures were designed to discard empty segments, as well as segments with less than one second of activity.

V. Speech/Music classification evaluation

CNN performances are estimated on SMC corpora and compared to baseline machine learning algorithms provided in `Scikit-learn` toolbox [17]: Support Vector Machines (SVM) with RBF kernel, and Gaussian Mixture Models (GMM) with diagonal covariance matrices and varying number of gaussians (32 to 512). Models are evaluated using input vectors corresponding to concatenated MFC frames, with contextual length varying between 1 (no context) and 50 (context of 500 milliseconds). Best results are obtained using the highest contextual length, which was limited to 50 in order to allow acceptable temporal resolution in following segmentation tasks (10 milliseconds time step and 500 milliseconds context).

Effectiveness is reported at the *frame level* and at the *media level*. Frame-level decisions are based on the raw instantaneous predictions of SMC models obtained for each input vector (500 milliseconds context). The frame-level decision correspond to the class (speech or music) with highest probability. Media-level decisions are based on the products across all frame-level probabilities per class obtained for an audio excerpt (100 frame-level predictions per second). Best instantaneous frame-level estimates does not necessarily produce best media-level estimates, nor best segmentations. SMC performances reported are those associated to the models achieving the best media-level performances, which is more representative of the final use-case.

Models are compared using a 5-fold cross-validation process. Effectiveness is described using speech, music and mean recalls, providing a description of the most frequent classification errors. MUSAN corpus evaluations are carried out using constraints on cross-validation, in order to group subtracks corresponding to a same performer in the same folds. These constraints avoid testing a model on a track produced by a performer found in the model's training set. Similar constraints are applied for segments obtained from the same audio-book or us-government track.

TABLE I. Speech/Music classification results obtained on GTZAN corpus [14] using 120 30-seconds long excerpts

Algorithm	*Frame-level Recall*			*Media-level Recall*		
	Speech	Music	Mean	Speech	Music	Mean
GMM	89,18	84,05	86,61	100	100	100
SVM	95,64	91,06	93,35	100	100	100
CNN	98,16	96,25	**97,21**	100	100	100

TABLE II. Speech/Music classification results obtained on Scheirer-Slaney corpus [15] using 240 15-seconds excerpts

Algorithm	*Frame-level Recall*			*Media-level Recall*		
	Speech	Music	Mean	Speech	Music	Mean
GMM	85,28	81,03	83,15	99,29	98,00	98,64
SVM	93,01	89,62	91,32	98,54	98,00	98,27
CNN	93,79	91,79	**92,79**	100	98,00	**99,00**

TABLE III. Speech/Music classification results obtained on MUSAN corpus [16] using 1948 3-seconds long excerpts

Algorithm	*Frame-level Recall*			*Media-level Recall*		
	Speech	Music	Mean	Speech	Music	Mean
GMM	97,32	95,04	96,18	99,41	97,17	98,29
SVM	97,17	94,5	95,84	99,8	97,41	98,6
CNN	99,36	98,85	**99,11**	100	99,46	**99,73**

Tables I, II and III report the results obtained on GTZAN, Scheirer-Slaney and MUSAN corpora. All classification algorithms achieve 100% correct media-level classification rate on GTZAN corpora. This better media-level recall is probably due to the duration associated to GTZAN samples (30 seconds, instead of 15 in Scheirer-Slaney and 3 in MUSAN). For all other tasks, CNN models achieve better classification results than GMM and SVM, both at the frame and media level. For all of these methods, frame-level recall is higher for speech than for music, as several music frames are incorrectly predicted as speech. CNN predictions are associated to a perfect media-level speech recall score for all corpora. CNN produces only 2 classification errors on Scheirer-Slaney corpus, and 5 on MUSAN corpus. Manual analysis of these errors shows vocal singing excerpts without instrumental accompaniment, or singing style close to regular speech, that are predicted as speech. An additional error found in Scheirer-Slaney corpus corresponds to a pure music track incorrectly classified as speech. This is explained by the presence of 3 training samples containing the same music track with superposed speech, learned as speech excerpts. Frame-level results obtained on Scheirer-Slaney corpus are worse than those obtained on other corpora, since this corpus contains speech-over-music excerpts, harder to discriminate than pure speech or pure music segments. Best frame-level results are obtained on MUSAN corpus; this may be explained by the low variability observed within its speech material.

VI. Speech/music segmentation evaluation

This section reports the early results obtained on the SMS task, using the CNN architecture associated to the best SMC performances. CNN model was trained using GTZAN, Scheirer-Slaney and MUSAN corpora described in the last sections, corresponding to about 3 hours of annotated material. Raw frame-level music and speech probabilities are obtained using a step size of 10 milliseconds. Viterbi algorithm is used to infer the most likely state sequence (speech or music) from these raw estimates, allowing to increase the robustness of CNN's predictions.

MIREX 2015 speech/music detection training examples material [6] was used for this evaluation (the training examples material is different from MIREX evaluation material which is not public). It contains about 5 hours of radio streams, corresponding to 4 hours of music, 1 hour of speech, 7 minutes of speech-over-music, and 6 minutes of other phenomena (pauses, applause, etc.). It is split into 7 tracks associated

TABLE IV. SPEECH/MUSIC SEGMENTATION RESULTS OBTAINED ON MIREX 2015 TRAINING EXAMPLES MATERIAL

Radio Stream Genre	*Raw Frame-level Recall*			*Frame-level Recall after Viterbi post-processing*		
	Speech	Music	Mean	Speech	Music	Mean
Classical	91.30	97.66	94.48	100	100	100
Country	92.94	73.45	83.19	96.82	88.55	92.69
Ethnic	95.15	71.49	83.32	99.18	79.26	89.22
Irish	95.08	88.32	91.70	99.38	96.50	97.94

to specific musical genres. Two tracks contain instrumental classical music. Three files contain ethnological recordings, including shamanic singing and psalms. One contains country music and blues, including *a cappella* singing. Last track contains Irish folk music. Table IV reports frame-level results obtained for all of these genres. Results for frames not associated to speech or music are not reported. Perfect results are obtained for material associated to instrumental classical music, which is the easiest category, well represented in our training database. Errors found in ethnological recordings are mostly associated to shamanic psalms, not covered in our training set, which are detected as speech. Errors found in the country music material are mostly associated to *a cappella* singing detected as speech. This source of error is coherent with studies reporting similar acoustic properties in country singer's speech and singing [18].

VII. DISCUSSION AND FUTURE WORK

This study presents the use of a CNN architecture on several SMC tasks and on a SMS task. The proposed model shows clear advantages over SVM and GMM, both at the *frame level* and at the *media level*. All speech segments used in classification evaluations, with to durations between 3 and 30 seconds, are correctly classified. The proposed semi-supervised training procedure allows to obtain slightly better SMC results than fully supervised approaches on MUSAN corpus (not reported in last sections), resulting in a reduction of two media-level errors over 1948 samples. This trend needs to be confirmed using harder SMC evaluation protocols, as well as larger evaluation corpora.

Errors related to music excerpts, recognized as speech, are mostly associated to *a cappella* or predominant vocals, and singing styles close to regular speech (hip hop, shamanic psalms, country vocals, opera *recitative* style, etc.). Scheirer-Slaney corpus is the only classification corpus used in this study including speech-over-music samples, resulting in the lowest frame-level classification results. These findings suggest to constitute training and evaluation corpora containing these difficult musical genres, and to systematically integrate speech-over-music samples in our future evaluation procedures.

Work in progress covers the constitution of a representative music segment dataset, with annotated variations related to genre, singing style, performer, and track identification. The use of speech databases containing speaker identities, and speaking styles annotations is required to improve control over our next evaluations. This issue may be partly addressed using a 2290-speaker corpus realized at Ina [19]. The hardest issue is the design of models able to discriminate speech-over-music and music, especially for music having vocal acoustic properties close to regular speech (hip hop, eletro, etc.). This will be addressed through data augmentation strategies, here consisting in artificially superposing speech to music [20]. The last issue is the constitution of a music track segmentation corpus, including early challenging archive documents representative of the diversity of Ina's collections [21].

REFERENCES

[1] L. Ericsson, "Automatic speech/music discrimination in audio files," Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan, 2010.

[2] R.-L. Jimena, R. Hennequin, and M. Moussallam, "Detection and characterization of singing voice using deep neural networks," UPMC-Paris, 2015.

[3] Y. Lavner and D. Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," EURASIP Journal on Audio, Speech, and Music Processing, no. 1, 2009, p. 239892.

[4] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts." in Interspeech, vol. 9, 2009, pp. 2583–2586.

[5] D. Castán et al., "Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains," EURASIP Journal on Audio, Speech, and Music Processing, no. 1, 2015, p. 33.

[6] "Mirex 2015 music/speech classification and detection and challenge," visited on 2017-03-07. [Online]. Available: http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection

[7] T. Lidy, "Spectral convolutional neural network for music classification," in Music Information Retrieval Evaluation eXchange, 2015.

[8] J. Royo-Letelier, R. Hennequin, and M. Moussallam, "Mirex 2015 music/speech classification," Music Inform. Retrieval Evaluation eXchange (MIREX), 2015.

[9] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol. 3361, no. 10, 1995, p. 1995.

[10] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[11] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research, vol. 15, no. 1, 2014, pp. 1929–1958.

[12] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in Neural Networks: Tricks of the Trade. Springer, 2012, pp. 561–580.

[13] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[14] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Transactions on speech and audio processing, vol. 10, no. 5, 2002, pp. 293–302.

[15] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., vol. 2, 1997, pp. 1331–1334.

[16] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[18] R. E. Stone, T. F. Cleveland, and J. Sundberg, "Formant frequencies in country singers' speech and singing," Journal of Voice, vol. 13, no. 2, 1999, pp. 161–167.

[19] F. Salmon and F. Vallet, "An effortless way to create large-scale datasets for famous speakers." in Language Resources and Evaluation Conference, 2014, pp. 348–352.

[20] J. Razik, C. Sénac, D. Fohr, O. Mella, and N. Parlangeau-Vallès, "Comparison of two speech/music segmentation systems for audio indexing on the web," in Proc. Multi Conference on Systemics, Cybernetics and Informatics, 2003.

[21] D. Doukhan and J. Carrive, "Simple neural representations of speech for voice activity detection and speaker tracking in noisy archives," 4th International Conference on Statistical Language and Speech Processing, 2016.