HIERARCHICAL RESIDUAL-PYRAMIDAL MODEL FOR LARGE CONTEXT BASED MEDIA PRESENCE DETECTION

Qingming Tang[1], Ming Sun[2], Chieh-Chi Kao[2], Viktor Rozgic[2], Chao Wang[2]

[1]Toyota Technological Institute at Chicago
[2]Amazon Alexa
qmtang@ttic.edu, {mingsun,chiehchi,rozgicv,wngcha}@amazon.com

ABSTRACT

We study media presence detection, that is, learning to recognize if a sound segment (typically lasting for a few seconds) of a long recorded stream contains media (TV) sound. This problem is difficult because non-media sound sources can be quite diverse (e.g. human voicing, non-vocal sounds and non-human sounds), and the recorded sound can be a mixture of media and non-media sound.

Different from speech recognition, where the recognizer needs to detect local phonetic variation, the key features used to distinguish media and non-media sounds are non-local features. Motivated by this, we propose a hierarchical model to learn representation of each pre-chunked segment within a long recorded stream jointly, and encourage every local representation to be not sensitive to variations within each segment. We also further explore the effects of techniques including stream based normalization and iteratively imputing missing labels of training dataset. Experimental results indicate that our proposed contextual based methods are effective for media presence detection.

Index Terms— Media Presence Detection, Contextual Based Model, Invariant to local variants

1. INTRODUCTION

Media presence detection [1] refers to the task of recognizing if there is any media sound in a (typically short) sound snippet. Being able to distinguish media sound and human sound has significant values for many real-life applications and products. For example, to reduce media related false triggers for virtual assistants such as Amazon Alexa, Google Assistant and Apple Siri. It also helps with analyzing user behavior for media content consumption. This concrete topic has close connection to a few popular research fields, including audio/video scene understanding and audio event detection.

For audio scene understanding, there are a bunch of works in the literature addressing the problem of recognizing the background (e.g. in mall, park or restaurant) of the acoustic signals [2, 3, 4, 5]. [2] studies what kind of acoustic features would be useful for acoustic background understanding. Different machine learning models, like Hidden Markov Model (HMM) [6], Support Vector Machine (SVM) [7] and Neural Networks (more specifically, in combination of Long Short-Term Memory (LSTM) [8] and Convolutional Neural Network (CNN) [9]) are used in [3, 4, 5] respectively for background understanding. [3] uses Mel-Frequency Cepstral Coefficients(MFCCs) while [4, 5] also explore cepstral, energy and voicing features besides spectrogram. The biggest difference between media presence detection and these acoustic background understanding works is that, media presence detection focuses on determining the existance of media sounds in a robust way, independent of what kind of background it is. Video genre analysis/classification [10, 11] typically have the same goal of audio scene understanding, where audio is a useful second modality in order to enhance the performance.

Audio event detection [12, 13, 14] focuses on identifying occurrences of events of interest (e.g. glasses breaking, coughing, gun shoot etc), or generic user activities in the given audio streams. The major difference between media presence detection and audio event detection is that, audio event detection more focuses on distinguishing a concrete type of event from the background. For example, audio event detection works on recognizing "dog barking" from other sounds and also silence. While media presence detection works on distinguishing "real dog barking" and "recorded dog barking".

Similar to recent audio event detection/classification works [15, 16], we also use weakly labeled data (e.g. indicating media presence or not without specifying the concrete boundary) for our media presence detection task. Our input data contains long-duration streams, each has been chunked into weakly labeled short (e.g. 5 second) segments.

A piece of recorded audio consists of both audio-level characteristics (e.g. speaker identity, F0, dynamic range and band limiting) and local information (e.g phonetic content) [17]. Key characteristics to distinguish if a sound is from media (TV) source or non-media sources are features that are shared by the whole piece of audio instead of local variation. This motivates us to learn a representation

that ignores the small local variations. We use bidirectional pyramidal LSTM [18, 19] (e.g. stacked bidirectional LSTM [8] with subsequent layer subsamples hidden states) on top of a residual network to extract representations for each labeled snippet.

Existing approaches (e.g. [1]) for media presence detection are typically based on chunked short duration audio (e.g. 5-second or 10-second pre-chunked short audio segments), which do not model longer contextual information. However, utilizing (very large) contextual information can boost the confidence of prediction, considering that media is typically be on/off for a continuous long duration. Besides, seeing contextual information can help a model to realize the on/off switch of media presence events in a large time scale. We would use a uni-directional LSTM to learn the contextual representation (historical memory) for a given stream, and the contextual representation is used together with local representation for per segment media presence detection. Such kind of hierarchical model is much faster than directly modeling the whole stream with a RNN, and it will not easily suffer from gradient vanishing problem. In this paper, we explore two variants of model architectures that towards using contextual information.

2. RESIDUAL-PYRAMIDAL LOCAL MODEL

We use in-house collections of media sounds for experiments. The length of audio streams range from a few seconds up to 30 minutes. Each stream is chunked into non-overlapping 5-second audio segments, each labeled as media presence or not. There are in total 2059 streams which consists of 191604 labeled, and 29421 unlabeled 5−second segments. For each stream, we can retrieve the order of the segments. There are 311 and 310 streams for validation and test set respectively, which corresponds to 16765 and 16939 5−second segments, respectively. We extract log mel-filter bank energy (LFBE) features from audio, with a window of $25ms$ shifted at every $10ms$. As a result, a 498×20 (as in Fig 1a) feature matrix is extracted for each 5-second segment.

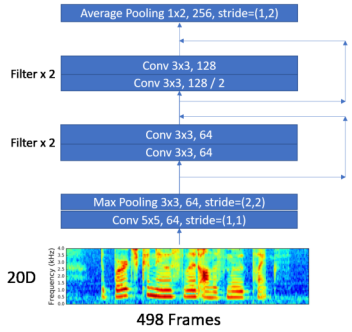
Our local model consists of a shallow Residual Network (ResNet) [20] and a stacked bidirectional pyramidal RNN with LSTM cells [18, 19]. The shallow ResNet (shown in Figure 1a) only consists of two residual blocks, preceded by one 5×5 convolutional and 3×3 pooling layer, followed by an average pooling layer. The ResNet transforms the $498 \times 20D$ input into a $125 \times 256D$ intermediate representation, reduce the length of temporal domain to roughly $\frac{1}{4}$ of the original. The representation is fed to stacked pyramid bidirectional RNNs, which is further followed by an average pooling layer and a softmax layer. The RNN cell size we used is 256 per direction. The output of the softmax layer is a real number between 0 and 1. The larger the value, the higher chance that the input 5−second audio segment (partially) contains media source sound. A pyramidal layer subsamples its input layer by a fixed scaling factor, which is equivalent to enforce segmental structure to hidden states of a layer without subsampling. Thus, stacking a few pyramid layers (with reasonable subsampling rate) would enforce the local model to ignore those very short term variations. This matches our motivation that features for media presence detection task should be “signal level" rather than “local/phonetic level". Note that, there are a few variations (e.g. skipping or concatenation) of pyramidal layer as shonw in [18], we adopt the skipping one. We use 3 pyramidal layers in our local model (downsample rate $\frac{1}{3}$).

We also applied a few regularization techniques including recently proposed recurrent dropout [21, 22], that is dropout applied to recurrent connections in RNNs. There is long argument that recurrent dropout would hurt performance of RNN based models [23] presumably because it hurts the precious historical memory of RNN. However, our experiments show that, recurrent dropout is capable to improve the performance of media presence detection task. We hypothesize this is because it makes the model more robust to very short term variations. For non-recurrent dropout, we selected from $\{0.2, 0.3, 0.4, 0.5\}$, while $\{0.0, 0.05, 0.15, 0.2\}$ recurrent dropout.

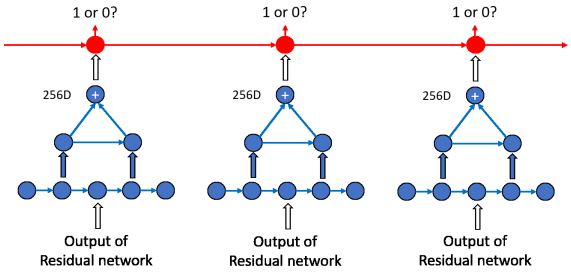
3. CONTEXT BASED PREDICTION

Each recorded stream might have very different recording environment and specific channel effect. One quick solution to address such per-stream difference is to remove the per-stream mean and variance for per-frame 20 dimensional LBFE feature vectors. As shown in the experimental study part (section 5), stream-based normalization is very helpful for local model presented in section 2. Actually, stream-based normalization is a direct way to utilize the contextual information.

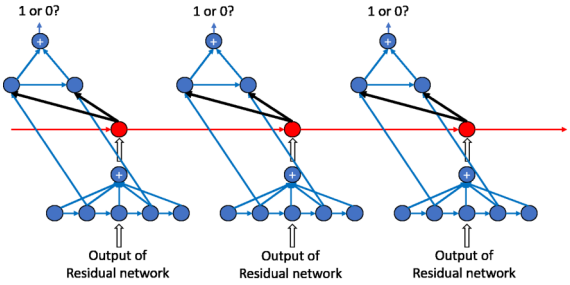
We thus move to contextual model, which aims at summarizing very broad contextual information using recurrent neural networks. Basically, our contextual model works as an ensemble of local models described in section 2. A unidirectional LSTM is used to chronologicaly process the representation from the local model's topmost (or intermediate) bidirectional pyramidal RNN layer, for each 5 second chunk. Such kind of model gathers together the long-range historical memory and local representation towards media presence prediction. There are several motivations on choosing such a hierarchical model rather than a flat RNN on modeling a very long stream. Firstly, the original stream can be too long, and a RNN directly running on (up to) 30min stream can be very time consuming. Secondly, a recurrent model can suffer from gradient vanishing problem (even using LSTM cell) if the input sequence is too long [24]. Finally, hierarchical model is a natural fit to our stream data consists of pre-chucked weakly labeled short segments.



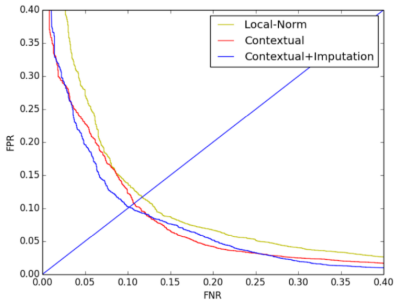
(a) Shallow residual network. A local model consists of shallow residual network plus pyramidal RNN (blue part in figure 1b and 1c)



(b) Contextual model I, a hierarchical model with representation from local model as input to contextual RNN. The blue part indicates the pyramidal part of local model, the "+" sign indicates average pooling layer and the red part indicates the contextual RNN.



(c) Contextual model II, a hierarchical model that learns representation of each local audio snippet conditioned on contextual representation. The intermediate layer of local models (blue) are averaged to be input to the contextual model (red part), and the output of contextual model is further fed back to be input of higher layer of local model.



(d) FPR-FNP curves for local model with stream-based normalization, contextual model II and contextual model II with imputation on test set.

Fig. 1: Local Model, Contextual Models and FPR-FNR curve

Each snippet is modeled by our residual-pyramidal architecture, while media on/off switch is modeled by contextual RNN. We use 2 pyramidal layers for local model module (in section 2 we use 3) within contextual models.

3.1. Model One

We present our first contextual model, which uses local model (without the softmax layer) as feature extractor to get a $256D$ per segment representation. These representations generated by average pooling layer (blue circle with "+" sign as shown in Figure 1b) are the input to a uni-directional RNN (LSTM cell). The hidden state of the RNN (red circle shown in Figure 1b) then has memory of history when making prediction via a softmax layer.

3.2. Model Two

In contextual model I (Figure 1b), the historical memory is included in the final representation used for prediction (e.g. red color circle in the figure). However, historical memory contains high-level and much richer information than predicting media presence or not for current segment. In this new model, we use two vectors to store historical memory (contextual information) and information more closely related to prediction respectively, as shown in Figure 1c. In contextual model I, the output from average pooling layer of figure 1a is the input to the contextual recurrent layer. However, in this model II, we feed representation of intermediate pyramidal layers from local module as the input to the contextual recurrent layer, which is a big difference compared to contextual model I. The output of the contextual recurrent layer, is further fed back to last pyramidal layer of local module followed by average pooling layer as shown in Figure 1c.

As we mentioned in section 2, labels are partially missing in training data. To overcome this problem, we add masks to the final per segment output of both contextual I and II, such that we only calculate loss for 5−second segments that have been annotated. In this way, those segments without annotation only contribute to learning historical memory. Actually, we can predict the labels of those training samples that are not annotated, and then use these "pseudo labels" to better calculate the loss in further training steps. Pseudo labels can be iteratively updated as training goes on. Our observation shows that such kind of imputation can boost the performance as shown in experimental section (section 5)

Table 1: Overall results using Equal Error Rate (EER), Accuracy (Acc), AUC, Recall at EER and Precision at EER. **D**=**Dev**, **T**=**Test**, ↑ (↓) means larger (smaller) is better. Contextual-II-Imp means Contextual-II with missing label imputation.

Methods	D EER(↓)	D Acc(↑)	T EER(↓)	T Acc(↑)	T AUC(↑)	T F1(↑)	T Recall(↑)	T Precision(↑)
Local	12.7	93.8	14.0	93.7	94.1	68.7	64.0	74.1
Local-Norm	11.0	94.0	11.3	94.2	95.5	68.8	58.5	**82.7**
Contextual-I	10.8	95.7	10.9	94.8	93.4	74.8	72.4	77.3
Contextual-II	**10.0**	**96.1**	10.7	94.3	95.8	**77.1**	74.9	79.1
Contextual-II-Imp	10.2	95.8	**10.1**	**95.2**	**96.1**	76.7	**75.3**	78.1

4. RELATED WORK

Media presence detection also has close connection to a few research topics, including Audio content representation and classification, Source separation and Spoofing detection.

For audio content representation and classification, existing works [25] typically focus on representing the audio towards storage and retrieval, thus typically do not consider the difficulty caused by mixture of sources. However, our daily life recorded stream typically consists of multiple sources of sound. Also, deep learning techniques are not (widely) used for the works listed in this field.

Source (signal) separation [26, 27] focuses on recovering the original signal from a mixture of sound, and spoofing detection [28] typically focuses on distinguish adversarial examples of specific users. Compare to the two fields, media presence detection is interested in distinguishing general in-person speaking and media sound.

5. EXPERIMENTS

We show the experimental results in this section. We compare different models using 5 different measurements. We first consider **Equal Error Rate (EER)**. This is the error rate when a selected threshold leads to equal false positive rate (**FPR**) and false negative rate (**FNR**). This is the major metric we would refer to on evaluating our methods. We would also check **Accuracy (Acc)** (when threshold equals 0.5), **Area under curve (AUC)** of Precision-Recall curve, **F1** score at EER, **Recall** at EER and **Precision** at EER.

We test our local model (without using stream based normalization) on the same training/dev/test set used in the paper [1] (but with single channel input), and we got 13.0%, 77.0%, 87.0% and 69.0% in terms of EER, F1, Recall and Precision. The results is better/comparable to the multi-channel model in paper [1]. The benefit presumably comes from the facts that our local model is seeking segmental-level invariant representation, and also the slightly deeper network. Thus, we think this local model itself can serve as a strong baseline.

Then, we test local model, local model with stream based normalization (local-norm), contextual model I (context-I), contextual model II (context-II) and contextual model II with imputation technique (contextual-II-Imp) using our in-house collection data. The overall results shown in Table 1 match our expectation that using contextual information is helpful. Basically, per stream normalization can also be interpreted as a way to utilize the contextual information, which significantly improves the performance over local model (local-norm vs local). Contextual model I learns a slightly better way to use contextual information comparing to per stream normalization according to row 3 and row 2 of Table 1. Contextual model II improves over contextual model I by storing historical information and local information separately.

Imputation based learning improves the performance in some metrics, but not all. However, as the number of missing labels are not huge, the observation is actually not conclusive here. We expect more stable improvement of imputation based learning, if we have more missing labels or if the label quality is not good.

Finally, we plot the FNR-FPR curve of three models local-norm, contextual-II and contextual-II with imputation on test set, as shown in Figure 1d. As show in the figure, contextual model II with/without imputation is consistently better than local model. Contextual model II with imputation is significantly better than contextual-II without imputation when FNR is relatively small (e.g. $0.05-0.1$).

6. CONCLUSION

We propose a residual-pyramidal hierarchical model architecture to capture contextual consistent, segment-level representation that is invariant to local (e.g. 5 second level) variances for media presence detection. Our experimental results show the benefits of our proposed model in the media presence detection task. We also study the effect of recurrent dropout and stream-based normalization to media presence detection. Our proposed techniques can be generalized to broader fields, such as acoustic event detection and spoofing detection. We have two follow-up directions based on this work. One is to explore structured learning for this task, e.g. we can apply conditional random field (CRF) on top of our existing contextual model. The other is to utilize latent variable models, e.g. treating the missing labels and label quality as latent variables, and use Expectation Maximization (EM) or Variational Auto-encoder (VAE) based methods to solve the problem.

7. REFERENCES

[1] Constantinos Papayiannis, Justice Amoh, Viktor Rozgic, Shiva Sundaram, and Chao Wang, "Detecting media sound presence in acoustic scenes," *Proc. Interspeech 2018*, pp. 1363–1367, 2018.

[2] Martin Cooke, Guy J Brown, Malcolm Crawford, and Phil Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, 1993.

[3] Brian Clarkson, Nitin Sawhney, and Alex Pentland, "Auditory context awareness via wearable computing," *Energy*, vol. 400, no. 600, pp. 20, 1998.

[4] Jurgen T Geiger, Bjorn Schuller, and Gerhard Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[5] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[6] Lawrence R Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] Christopher JC Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

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

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] Kyu-Phil Han, Young-Sik Park, Seong-Gyu Jeon, Gwang-Choon Lee, and Yeong-Ho Ha, "Genre classification system of tv sound signals based on a spectrogram analysis," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 1, pp. 33–42, 1998.

[11] Matthew Roach and John S Mason, "Classification of video genre using audio," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[12] Weiran Wang, Chieh-chi Kao, and Chao Wang, "A simple model for detection of rare sound events," *Proc. Interspeech 2018*, pp. 1344–1348, 2018.

[13] Chieh-Chi Kao, Weiran Wang, Ming Sun, and Chao Wang, "R-crnn: Region-based convolutional recurrent neural network for audio event detection," *Proc. Interspeech 2018*, pp. 1358–1362, 2018.

[14] Bowen Shi, Ming Sun, Chieh-Chi Kao, Viktor Rozgic, Spyros Matsoukas, and Chao Wang, "Compression of acoustic event detection models with low-rank matrix factorization and quantization training," *NeurIPS 2018 workshop on Compact Deep Neural Networks with industrial applications*, 2018.

[15] Anurag Kumar and Bhiksha Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[16] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 641–645.

[17] Wei-Ning Hsu, Yu Zhang, and James Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[18] Liang Lu, Lingpeng Kong, Chris Dyer, Noah A Smith, and Steve Renals, "Segmental recurrent neural networks for end-to-end speech recognition," *Interspeech 2016*, pp. 385–389, 2016.

[19] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Yarin Gal and Zoubin Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019–1027.

[22] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth, "Recurrent dropout without memory loss," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1757–1766.

[23] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[24] Asier Mujika, Florian Meier, and Angelika Steger, "Fast-slow recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 5915–5924.

[25] Stan Z Li, "Content-based audio classification and retrieval using the nearest feature line method," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 619–625, 2000.

[26] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[27] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[28] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–5.