DISCRIMINATE NATURAL VERSUS LOUDSPEAKER EMITTED SPEECH

Thanh-Ha Le[1,2], *Philippe Gilberton*[1] *and Ngoc Q. K. Duong*[1]

[1]Technicolor Research & Innovation, France
[2] Eurecom, France
leth@eurecom.fr, {philippe.gilberton, quang-khanh-ngoc.duong}@technicolor.com

ABSTRACT

In this work, we address a novel, but potentially emerging, problem of discriminating the natural human voices and those played back by any kind of audio devices in the context of interactions with in-house voice user interface. The tackled problem may find relevant applications in (1) the far-field voice interactions of vocal interfaces such as Amazon Echo, Google Home, Facebook Portal, *etc*, and (2) the replay spoofing attack detection. The detection of loudspeaker emitted speech will help avoiding false wake-ups or unintended interactions with the devices in the first application, while eliminating attacks involve the replay of recordings collected from enrolled speakers in the second one. At first we collect a real-world dataset under well-controlled conditions containing two classes: recorded speeches directly spoken by numerous people (considered as the *natural* speech), and recorded speeches played back from various loudspeakers (considered as the loudspeaker emitted speech). Then from this dataset, we build prediction models based on Deep Neural Network (DNN) for which different combination of audio features have been considered. Experiment results confirm the feasibility of the task where the combination of audio embeddings extracted from SoundNet and VGGish network yields the classification accuracy up to about 90%.

Index Terms— Natural versus device emitted speech classification, voice interaction, audio features, deep neural network (DNN).

1. INTRODUCTION

Vocal interface devices such as the Apple Home Pod[1], Google Home[2], Amazon Echo[3], and Facebook Portal[4] *etc.,* are becoming incredibly popular at home nowadays. These smart devices use primarily the automatic speech recognition (ASR) function to recognize user's command and spoken context. When deploying in real acoustic environments, the device will actually capture speech from both the target users and those emitted from loudspeaker devices such as radio or television. To tackle this issue, existing approach was building device-directed speech detection systems [1, 2, 3, 4, 5, 6] so as only the detected target speech is taken into account for the interaction. To improve the detection accuracy, speaker identification function can be added where only speech from a pre-defined list of users is processed. As examples, Google Home supports maximum 6 users in one device. Amazon Echo works with multiple user accounts, but they must be switched manually by explicitly asking Alexa to do it. To support further the multi-users vocal interface, we propose to investigate in this paper another capability to distinguish between the speech directly spoken by users, we refer hereafter as *natural speech*, from those emitted by loudspeaker devices. This considered problem may greatly help the vocal interface devices to avoid annoying false wake-ups, unintended interaction, or irrelevant processing of recorded sounds.

Another important application of the considered problem is to protect automatic speaker verification (ASV) from replay spoofing attacks (RSA) where the hackers replay recordings collected from enrolled speakers in order to provoke false alarms. The RSA is known to be very challenging as recordings made with high quality hardware may be close to indistinguishable from genuine speech signals [7, 8, 9, 10]. The considered problem does the same task with the ASVspoof 2017 Challenge[10]. However, the ASVspoof 2017 dataset [11] was not recorded under a fully well-controlled acoustic environments which didn't fit well with our in-house context. That is the reason why we just select the ASVspoof 2017 dataset as a part of our dataset.

The considered classification problem is challenging due to the variations of the natural speech from different people, the hardware devices (microphones for recording, loudspeakers to play back), and the far-field acoustic environments. Thus, in this work we first devote our effort in collecting a large-scale dataset with significant diversification of hardware devices as well as the number of recorded users in different environments. To enrich the dataset we inherit the dataset available from the ASVspoof 2017 Challenge [11]. Then we investigate different audio features extractors like (*i.e.,* Mel-frequency cepstral coefficients (MFCC), constant Q cepstral coefficients (CQCC), embeddings from SoundNet and VG-

---

[1]https://www.apple.com/homepod/
[2]https://store.google.com/product/google_home
[3]https://www.amazon.com/echo
[4]https://portal.facebook.com/

Gish) for the prediction model.

The rest of the paper is organized as follows. Section 2 presents the dataset collection. Section 3 focuses on building DNN-based prediction models. Experimental results are discussed in Section 4. Finally, we conclude in Section 5.

2. DATASET

In this section we will first describe the inherited dataset from the ASVspoof 2017 Challenge. We then present our own data collection where acoustic environments are well-controlled. Finally we describe the data split and augmentation for training and testing the prediction model.

2.1. Anti spoofing replay attack dataset

The ASVspoof 2017 dataset is a collection of bona fide and spoofed utterances. The bona fide are a subset of RedDots corpus [12]. Spoofed utterances are recorded by replaying and recording bona fide utterances using different devices and acoustic environments. The replay spoofing scenario and replay configurations are particularly described in [11]. The ASVspoof 2017 dataset is recorded in a diverse range of replay configurations including 26 playback devices, 25 recording devices and 26 acoustic environments. In our work, we use the ASVspoof 2017 version 2.0 dataset which was released from training, development and evaluation subsets. Since ASVspoof 2017 is considered as challenging dataset, the number of samples and replay configurations in the evaluation subset is larger than training and development subsets. We repartioned the ASVspoof 2017 version 2.0 dataset into three subsets: training, validation and testing subsets which consist of 60%, 20% and 20% samples, respectively.

2.2. In-house dataset

We first recorded natural speech in two languages (English and French) from ten people. We then performed denoising on the recorded natural speech signals and played them back by loudspeakers to record the loudspeaker emitted samples. The details of recording configurations are summarized as follows.

Acoustic environments: we recorded the dataset in two rooms: small size of 3 m^2 and medium size of 15 m^2. In each room, ambient noise was mainly produced by air conditioning systems, but sometimes by people walking along the corridor.

Distance from speakers/loudspeakers to the microphones: In each room, we recorded natural voices with two distances (*i.e.,* one meter and three meters) from people to the microphones. Various distances in different indoor environments may diversify greatly our dataset so that it can minimize the classifier to learn on specific acoustic conditions for the two classes.
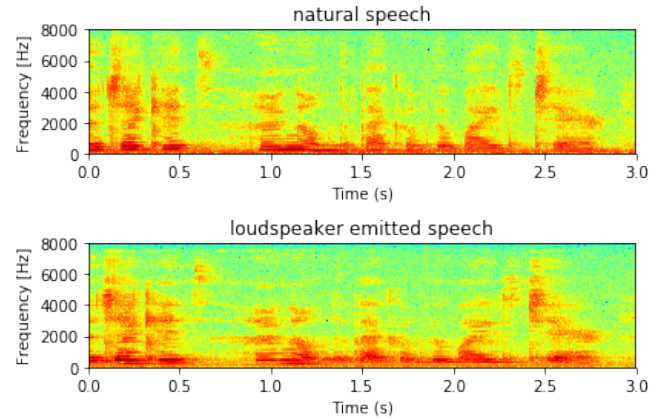

Fig. 1. Spectrogram of the natural versus loudspeaker emitted speech of sentence 'The jacket hung on the back of the wide chair'.

Devices: To avoid the classifier learning on a specific device, we recorded all speech signals with two different microphones based on two different technologies (a high end one based on electret sensor and a low cost one based on MEMS sensor). For the recording of the emitted speeches, we replayed the natural speech signals by using three different loudspeakers, among which two are from a TV set as they corresponded the most to the actual source of background speech in a home environment.

Languages: To increase the variation of language phonemes, the dataset was built using French and English languages: each person read an English paragraph, followed by a French paragraph. The duration for reading each paragraph varied from 81 seconds to 206 seconds.

Overall, the dataset consisted in speech recordings upon two acoustic environments, with two different ranges between speaker/or loudspeaker and the microphone, two microphones, three loudspeakers, two languages, and ten people, resulting in a total of 480 different configurations. All samples were mono channel, recorded with a sampling rate of 16000Hz, 16 bit PCM and encapsulated into wav format. Note that, in order to avoid the machine learning to build a model that could be biased on the recording conditions between two classes (instead of learning the actual differences of spectral patterns between the natural and the loudspeaker emitted recordings), we positioned the loudspeaker at the same location as it was performed during the natural speech recording of the person. Thus the collected dataset could be more challenging than the ASVspoof 2017 dataset as each pair of samples in the two classes in the former was recorded in the identical acoustic environment. As an example, Fig. 1 shows the spectrograms of a pair of recordings in such two classes, which look very similar.

We segmented the recordings into 1 s length audio segments with 50% overlap, and partitioned all the obtained segments into 3 subsets: training, validation and testing. The number of audio samples in each set is summarized in Table 1. Note that for a better model generalization, all these subsets were disjointed in terms of speakers.

#seg in training		#seg in validation		#seg in testing	
Natural	Emitted	Natural	Emitted	Natural	Emitted
5053	9832	5163	10263	8390	16527

Table 1. Number of audio samples (natural and loudspeaker emitted segments) in the training, validation, and testing set of the In-house dataset.

2.3. Data organization and pre-processing

After segmenting the recordings into 1 s segments with 50% overlap, we normalized the segments' energy to the same level. The new dataset was obtained by merging the ASVspoof 2017 dataset and the In-house dataset. We merged together the training set, validation set and testing set of these two dataset. The final number audio samples in each set of the overall dataset is shown in Table 2. As it can be seen, the dataset is not perfectly balanced as the number of loudpeaker emitted segments is about three times bigger than the number of natural speech segments in each subset.

#seg in training		#seg in validation		#seg in testing	
Natural	Emitted	Natural	Emitted	Natural	Emitted
15068	51186	8386	24319	11409	30270

Table 2. Total number of audio samples (natural and loudspeaker emitted segments) in the training, validation, and testing set of the overall dataset.

3. PREDICTION MODEL

Prior to building the prediction model, we first extracted a range of state-of-the-art audio features for each 1 s audio samples in the dataset. The details of the extracted features will be presented in Section 3.1. We then investigated the use of multi-layer perception (MLP) and support vector machine (SVM) as classifier. After using grid search to optimize the parameters for both MLP and SVM, we found that MLP gave slightly better prediction accuracy than SVM. Thus we will focus preferably on MLP that will be our choice of classifier for which we will report its results in the experiment.

3.1. Audio feature extraction

MFCC: as a well-known feature extractor for automatic speech recognition, in our investigation we extracted 12 MFCC coefficients for each overlapping audio frame of 25 miliseconds (hope size is 10 miliseconds). As a result, we obtained a 1212-dimensional MFCC feature for each 1 s audio sample in the dataset. To make the feature dimension comparable to VGGish and SoundNet features, we further performed a Principal Component Analysis (PCA) to reduce the MFCC feature dimension to 128. The 128-dimensional feature may be more relevant for training the model given the moderate size of the dataset. In the experiment, we will report prediction results obtained with both the 1212-dimensional and 128-dimensional MFCC features.

CQCC: this feature uses constant Q transform to offer greater frequency resolution in lower frequencies and greater time resolution in higher frequencies. As loudspeaker devices often make more signal distortion at low and high frequency ranges, we expect that the CQCC can distinguish such differences between the natural versus loudspeaker emitted speech, and thus potentially improve the prediction result. Note that CQCC was also used as main feature in the baseline system of the ASVspoof 2017 challenge [13]. In our experiment, the constant Q transform is applied with a maximum frequency of 8000Hz, the minimum frequency is set to 15Hz, the number of bins per octaves is 96 and the sampling period is 16. Each 1 s audio sample in the dataset results in a 1404-dimensional CQCC feature. Similarly to the MFCC, we further used PCA to reduce the feature dimension to 128 to be balance with other DNN-based features and to be more relevant with the actual size of the dataset.

SoundNet feature: SoundNet was designed to learn rich natural sound representations from large amounts of unlabeled sound data collected in the wild [14]. It was targeted to transfer discriminative visual knowledge from convolutional neural network (CNN)-based visual recognition models into the sound modality using 2,000,000 unlabeled videos. Soundnets embedding features were reported to obtain the state-of-the-art result for acoustic scene classification task on three benchmark datasets: DCASE Challenge [15, 16], ESC-50 and ESC-10 [17]. In our experiment, we used the pre-trained SoundNet based on 8 convolutional layers and we extracted embeddings from different layers for testing: conv4, conv5, conv6, and conv7. The best classification result (reported in the experiment section) was obtained with the 512-dimensional embedding at the layer conv5.

VGGish feature: Released recently by Google[5], VGGish is a deep convolutional neural network trained on Audioset dataset [18]. Audioset is a dataset of over two million human-labeled 10-second YouTube video soundtracks, with labels of more than 600 audio event classes [19]. VGGish is a variant of the well-known VGG model designed for image classification. The input size was 96x64 for log mel spectrogram audio inputs. The last group of convolutional and maxpool layers is dropped, so that VIGGish has four groups of convolution/maxpool layers instead of five as the VGG11. At

---

[5]https://github.com/tensorflow/models/tree/master/research/audioset

the end of the network, a 128-wide fully connected layer is used and this layer acts as a compact embedding layer. In our experiment, we extract this 128-dimensional VGGish embedding for each 1-s audio sample in the dataset.

Combined features: Since SoundNet was trained with unlabeled data and achieved state-of-the-art results in acoustic scene detection benchmark datasets, and VGGish was trained with labeled audio object classes, we concatenated embeddings extracted from those two neural networks with expectation that the new features can benefit from representing both information about audio scenes and audio objects. We also investigated the combination (by concatenation also) of all above mentioned features after balancing their dimension (*i.e.,* to 128 for MFCC and CQCC).

3.2. Model training

We tested two classification models: SVM and MLP for each type of features described in the Section 3.1. During the training, we varied the model parameters with grid search, *i.e.,* kernel types and their corresponding hyper-parameters for SVM; number of layers, number of neurons in each layer, batch size, step size, optimizer, *etc.,* for the MLP. We also performed the 5-fold cross validation to choose the best performing model in term of the classification accuracy.

4. RESULT

The best results offered by each type of audio feature were obtained by averaging the corresponding classification accuracy in each fold and shown in Table 3. Note that we do not show the results given by the SVM as in all cases MLP performs better. We also report the results of a Gaussian mixture model (GMM) classifier that was trained with CQCC feature as baseline model of the ASVspoof 2017 Challenge[10].

Compared to the baseline model of ASVspoof 2017 Challenge, all MLP classifiers performed better. As expected, 128 dimensional CQCC outperformed MFCC in the validation and test sets as CQCC was designed to better capture acoustic variations at low and high frequencies. On the other hand, DNN-based features SoundNet and VGGish performed better than the signal processing based features MFCC and CQCC as DNN trained on numerous data with different recording quality may capture richer information about audio signal, especially at low and high frequency ranges where the difference between the two sound classes may be more visible. This was also in line with current observations in the acoustic scene classification[6] [14] and event detection tasks[7]. VGGish feature, which was released recently by Google, offers the best testing accuracy of 88.04%. Overall the combination

---

[6]http://dcase.community/challenge2018/task-acoustic-scene-classification

[7]http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio

of VGGish and SoundNet features offer the best classification performance for both the training (90.04%), validation (88.54%), and testing (90.43%) sets. We also observed that a further combination of VGGish and SoundNet with MFCC and CQCC did not bring benefit as there might be acoustic redundancy in such combination.

The best MLP model was trained with combination of VGGish and SoundNet features has one fully connected layer with 100 neurons and softmax is used in the last layer. The learning rate is 0.00005, batch size is 5000 and the optimizer is Adam.

Audio features	#dim.	Accuracy %		
		Training	Validation	Testing
MFCC	1212	83.10	83.84	84.05
MFCC	128	89.84	84.93	84.79
CQCC	1404	87.58	84.64	84.95
CQCC	128	86.59	85.53	85.29
SoundNet	512	89.45	86.00	86.87
VGGish	128	88.58	86.05	88.04
VGGish+SoundNet	**640**	**90.40**	**88.54**	**90.43**
VGGish+SoundNet + CQCC+MFCC	512	89.77	86.36	88.68
Baseline model of ASVspoof 2017	2223	76.31	76.30	72.63

Table 3. Classification results on the training, validation, and test set obtained by MLP classifier with the use of different audio features.

5. CONCLUSION

In this work, we discussed a novel problem of detecting whether a speech recording is *natural*, *i.e.,* directly spoken by a person, or emitted from a loudspeaker device. Reliable detection model can find usage in emerging vocal interfaces and anti-spoofing application. We designed a large-scale dataset under well-controlled conditions to better fit with home environment and to allow us to train DNN-based prediction models. The best system with 90.43% test accuracy was obtained by using the combination of VGGish and SoundNet embeddings as audio feature extractor and MLP as classifier. This result confirms the feasibility of the task despite such a challenging classification problem. Future work may be devoted in enlarging the dataset with more users and more variety of loudspeaker/microphone devices to confirm the results. Fine-tuning VGGish or SoundNet so as the prediction system is learned end-to-end directly from input audio recordings is also a potential way to improve the prediction accuracy, especially with the enlarged dataset. Finally, a more analysis about what is the essence for discriminating the two classes of speech signals is planned.

6. REFERENCES

[1] S. Harish Mallidi, R. Maas, K. Goehner, A. Rastrow, S. Matsoukas, and B. Hoffmeister, "Device-directed utterance detection," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2018.

[2] E. Shriberg, A. Stolcke, D. Hakkani-Tr, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2012, pp. 334–337.

[3] D. Heger J. Ijsselmuiden R. Stiefelhagen D. Reich, F. Putze and T. Schultz, "A real-time speech command detector for a smart control room," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 2641–2644.

[4] T. Yamagata, T. Takiguchi, and Y. Ariki, "System request detection in human conversation based on multiresolution gabor wavelet features," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2009, pp. 256–259.

[5] H. Lee, A. Stolcke, and E. Shriberg, "Using out-of domain data for lexical addressee detection in human-human-computer dialog," in *Proc. NAACL*, 2013, pp. 221–229.

[6] D. Wang, D. Hakkani-Tur, and G. Tur, "Understanding computer-directed utterances in multi-user dialog systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 8377–8381.

[7] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2014, pp. 157–168.

[8] Z. Wu, S. Gao, E. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2014, pp. 1–5.

[9] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143–153, 2015.

[10] T. Kinnunen, Md Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 2–6.

[11] M.; Sahidullah Md; Evans N.; Kinnunen T.; Lee K.; Yamagishi J. Delgado, H.; Todisco, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *The Speaker and Language Recognition Workshop (ODYSSEY)*, 2018.

[12] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, Md. J. Alam, A. Swart, and J. Perez, "The reddots data collection for speaker recognition," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 2996–3000.

[13] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Proc. The Speaker and Language Recognition Workshop (ODYSSEY)*, 2016, pp. 283–290.

[14] C. Vondrick Y. Aytar and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. Neural Information Processing Systems (NIPS)*, 2016.

[15] E. Benetos M. Lagrange D. Stowell, D. Giannoulis and M. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[16] A. Mesaros, Toni Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Int. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[17] Karol J Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Multimedia (MM)*, 2015, pp. 1015–1018.

[18] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[19] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audioset: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.