Detection of Robocall and Spam Calls Using Acoustic Features of Incoming Voicemails

Benjamin Elizalde and Dimitra Emmanouilidou
Microsoft Research, Redmond, WA, USA
{benjaminm, dimitra.emmanouilidou}@microsoft.com

Index terms – spam calls; robocalls; feature selection

Abstract

Spam communications are organized attempts mainly aimed at marketing, at spreading false information or at deceiving the end recipient. Prerecorded messages called *Robocall*s are increasing every year, striking US residents with 46 billion calls in 2020. Carrier and telecommunication regulators have automated systems in place to identify unwanted calls using Call Detail Records, which include call origin or call duration information, but the actual audio content is often overlooked. We propose an audio-based spam call detection method that uses acoustic features of recorded voicemails to identify *Human* calls from *Robocall*s, and to identify *Spam* calls from *Non-Spam* calls for human callers. Results show that voiced and unvoiced audio content carry sufficient discriminatory information to distinguish between *Human* and *Robocall* but also between *Spam* and *Non-Spam* calls. Distinguishing between *Human* calls and *Robocall*s achieved 93% accuracy, compared to *Spam* vs *Non-Spam* achieving 83% accuracy. We expect that our automated approach can serve as an auxiliary tool, in combination with other call behavior statistics, to reduce the frequency of unwanted calls or fraudulent incidents.

1. INTRODUCTION

Unwanted or unsolicited calls represent a continuous threat for people around the world and have reached record numbers in recent years. Users worldwide received over 31 billion spam calls in 2020, up from 26 billion during the same period last year, and 17.7 billion the year prior, according to a recent data report [*]. There are various types of unsolicited calls, most commonly intended for broadcasting, telemarketing, scamming or extracting users' financial information, and can be carried on by a live person or by prerecorded messages like Robocalls. The US alone had 46 billion Robocalls in 2020. [†]

The diversity in the type of unwanted call messages makes it difficult to have a one-fits-all approach for detecting or preventing unsolicited calls. Current solutions rely on collaborations among different telecommunication sectors and governmental agencies, and on a variety of implemented protocols. Call Detail Record (CDR) is a data record produced by a telecommunication equipment exchange containing information on the details of the call, such as time, duration, call source or destination. CDR has been helpful to identify unwanted caller ID sources, however it has also raised privacy concerns due to storing user records. Other efforts include issuing of a digital certificate per caller ID, validated by telecommunication carriers.

[*] https://techcrunch.com/2020/12/07/spam-calls-grew-18-this-year-despite-the-global-pandemic/

[†] www.prnewswire.com/news-releases/americans-hit-by-just-under-46-billion-robocalls-in-2020-says-youmail-robocall-index-301215139.html

These protocols have contributed in a significant reduction of unwanted calls, though as spamming technology evolves, users are still getting deceived or even harmed.

Prior published work on spam detection took advantage of CDR and spammers' call pattern information on typical corpus sizes of 200-300 call incidents.[1] Amanian et al.[2] proposed to assign weight coefficients to call behavior and pattern features, like call duration or number of simultaneous calls, and then use Linear Discriminant Analysis to classify suspicious incoming calls. Authors in[3] proposed a novel deep learning-based approach to classify phone calls as Spam, based on the content of the generated call transcript. Authors in[4] proposed the use of acoustic features, such as tonality or spectral flatness, on 28 Spam Over IP Telephony (SPIT) calls, to detect identical or close to identical calls being received by multiple users. In this work, we are interested in a more generalized version of this problem, looking for specific audio-based features or acoustic patterns in a voice call that will help differentiate any *Spam* call from any *Non-Spam* call, without taking into account the call transcription or the similarity of identical outgoing calls.

We propose an audio-based spam detection method that uses acoustic features of recorded voicemails to identify *Human* calls from *Robocall*s, and to identify *Spam* calls from *Non-Spam* calls for human callers. Our approach does not rely on pairs or groups of identical calls, but it is rather an attempt to obtain more global features per category; furthermore, the extracted features do not rely on speech content, thus efficiently preserving user-privacy compared to prior art. To the best of our knowledge, this is the first attempt to explore these tasks with an audio-based methodology. Results show that acoustic features carry sufficient discriminatory information to make both tasks possible with high confidence. We expect that our automated approach can serve as an auxiliary tool, in combination with other call behavior statistics, to reduce the frequency of unwanted calls or fraudulent incidents.

2. METHODS

In this section we describe the collected voicemail corpus and available annotations, along with the feature extraction process and classification models.

A. DATA

Table 1: Overview of the voicemail corpus.

	Spam	*Non-Spam*	Other	(Male,Female)	Total
Human	118	129	39	(91,154)	286
Robocall	167	4	71	-	242
Uncertain	14	1	53	-	68
Total	299	134	163	-	596

Voicemails were extracted from the personal mobile devices of enrolled participants. The study imposed no restrictions as to the network carrier (GSM, CDMA, 4G, or 5G) and users were asked to remove voicemails they may not want to share. The majority of voicemails were in English. IRB approval was obtained for this study by the Microsoft Research Ethics Board. The dataset has not been made publicly available, respecting user privacy as these are personal phonecalls. In total, 596 voice calls were collected, with an average call duration of 37 secs ± 17 and median duration of 21 seconds. We will refer to the voicemail corpus as *calls* for the rest of the manuscript. For each one of the following attributes, one label was obtained per call, after enrolling three annotators for the task.

- *Human* call or not: Annotators had the option to choose between *Human*, *Robocall*, and *Uncertain* (unsure). This study will focus only on types *Human* and *Robocall* as indicated by the highlighted cells in Table 1; class type *Robocall* refers to calls from computer-generated voices.
- *Spam* call or not: Annotators were presented with the following choices: *Spam*, *Non-Spam*, *Ham*, and *Uncertain* (unsure), where *Ham* refers to requested or subscribed callbacks (for e.g. an announcement by the local utility company). In this study, we will only focus on the *Spam* and *Non-Spam* groups from human callers (of the *Human* category).
- Perceived speaker gender: Here, annotators were asked to indicate the perceived gender of the caller for Human calls. 154 of the *Human* calls were annotated as female 91 as male calls, and 2 were unknown. For reference, the corpus came from the mobile devices of 2 users that identified as he/him and 1 user that identified as she/her. The gender attribute was not used for classification purposes.

Table 1 provides an overview of the corpus. Speaker identity information was not available for this corpus, and data was randomly split into train and test sets. Recordings were down-sampled to 8000 Hz (the minimum sampling rate among all downloaded voicemails) and processed with a state-of-the-art Voice Activity Detector by Braun et al.,[5] for removing silent regions preceding or succeeding the call. Corrupted calls, calls shorter than 4 secs or without voice activity were discarded. For the task of distinguishing *Robocall*s from *Human* calls in Section 3.2, there are 285 *Human* and 242 *Robocall*s remaining files total, and we include all recordings irrespective of their *Spam*, *Non-Spam*, or Other label. For the task of distinguishing *Spam* from *Non-Spam* calls for Human callers in Section 3.3, we include calls in the set $\{$ *Human* $\cap$ *Spam* $\cap$ *Non-Spam* $\}$, where there are 128 *Non-Spam* and 118 *Spam* calls remaining after discarding corrupted data.

B. FEATURE EXTRACTION AND CLASSIFICATION

Audio-based spectro-temporal features were extracted for all calls on either the full recording or on a variable fixed-size segment. Two types of features were used: a) a 40 channel log-mel spectrogram calculated with 32 msec frames with 10 msec shift; and b) the open-source GeMAPSv01b OpenSmile set of 62 low level descriptors,[6] including spectral and temporal features and statistics from both voiced and unvoiced segments, such as speech pitch, speech formants, spectral slope and loudness, among others.

For classification, we split the data into two views: *Robocall* vs *Human* calls, and *Human Spam* vs *Human Non-Spam* calls. We used a statistical model, Support Vector Machines (SVM), with a linear (L-SVM) and an RBF kernel (K-SVM). Data was randomly split into $80-20\%$ for training and testing respectively and performance was calculated over 100 Monte Carlo runs; the margin of the SVM was unweighted according to the class importance to account for the small-scale corpus class imbalance.

An additional data driven approach was used, comprising of a 2D 2-block Convolutional Neural Network (CNN). The first CNN block has 32 channels, kernel size of 5 and stride of 1, followed by a Max-Pool layer; the second CNN block has 64 channels, kernel size of 3 and stride of 2 followed by a Max-Pool layer and dropout of 0.5, resulting in a 512-dimensional embedding; the end-classifier comprised of 3 fully connected layers of input dimensions $512, 2048, 128$ and a ReLu non-linearity was introduced after the second layer. We used a fixed batch size of 128, and used the validation set to select the best epoch, $e = 73$, with $e_{max} = 100$; the learning rate $lr = 0.002$ was optimized in $lr \in \{0.0001, 0.001, 0.005, 0.01, 0.02, 0.05, 0.1\}$. To train the CNN, data was split into $75 - 10 - 15\%$ for train, validation and test respectively, with 20 random network and data splits initializations. Besides the learning rate and the epoch selection, we did not tune any additional hyper-parameters. To the best of our knowledge this is the first published study of its kind and we chose to focus on the feasibility aspects of this work rather than its parametrization. For training, we only retain 10 sec of a call, ensuring fixed-length input features for the CNN. We augment the data by i) shifting the 10 sec window across the full recording and by ii) an additional 1X augmentation using SpecAugment[7]

with default parameters. The augmentation was used on the CNN approach to avoid overfitting of the learnt network weights.

3. RESULTS

Very little is publicly known in the area of unwanted voice calls and no public data is available, mainly to respect user privacy. We dedicate the following section to present statistical findings of the labeled data at hand, followed by results on the two main classification tasks.

A. FINDINGS ON THE ANNOTATIONS

Call duration statistics in this dataset reveal that the average call duration was 30 ± 27 sec long. Within the various call classes, Table 2 (left) shows that the average duration of *Human* calls were shorter by 10 sec compared to *Robocall*s; the average *Human Non-Spam* call was also found to be 10 sec shorter than the average *Human Spam* call. In our labelled data, we found that among all annotated *Non-Spam* calls (excluding *Uncertain* and *Ham* as mentioned in Section 2.1), 100% were *Human* calls. Looking into the *Spam* calls alone, 39% were annotated as *Human* calls and 56% were *Robocall*s. In a flip-side view, Table2 (right) shows the *Spam* and *Non-Spam* breakdown observed within the *Human* and *Robocall* classes. In this table, *Other* refers to either *Uncertain* or *Ham* annotations.

Table 2: Left: call duration statistics of this corpus. Right: percentage of Spam and Non-Spam calls found within the Human and Robocall annotations (left columns); speed of annotations (right column).

Call duration (sec)	μ	σ	$median$
Human	33	$\pm$ 25	29
Robocall	43	$\pm$ 25	33
Human Spam	37	$\pm$ 27	32
Human Non-Spam	27	$\pm$ 17	21

	Human class	*Robocall* class	An/t. Speed
Spam	48%	69%	6.2 sec
Non-Spam	52%	17%	7.7 sec
Other	-	14%	

We further looked at the speed of annotation, or in other words for how long did annotators need to listen to a call before making a decision; on average it took 6.6 sec for a reviewer to decide whether a call was *Spam* or *Non-Spam* with reviewers taking slightly longer to label a *Non-Spam* call (7.7 secs) compared to a *Spam* call (6.2 secs). We excluded outliers before calculating the average statistics.

Looking into the voice gender breakdown of the data, it appears that a *Human Spam* call was twice or more likely to come from a (perceived) female voice. 70% of *Human Spam* calls were labelled as female, and 30% as male voices. The gender ratio seems more equally distributed for the *Human Non-Spam* calls, were 56% were labelled as female and 44% as male calls.

B. FINDINGS ON *ROBOCALL* VS *HUMAN* CALLS

In this section we want to answer whether we can distinguish *Robocall* from *Human* calls using low-level audio-based features. For the classification of *Robocall* vs *Human* calls, we used the full duration of each call, with calls randomly split into 80%-20% for training and testing. We report unweighted (balanced) accuracy results over 100 Monte Carlo (MC) runs of a linear (L-) and an RBF (K-) SVM classifier with fixed parameters $\gamma = 1$ and $C = 1$ (no hyper-parameter tuning was performed); see Table 3, first three columns.

To illustrate the level of complexity for this classification problem, we explored best ways of selecting subsets of features. We used Recursive Feature Elimination and Forward and Backward Sequential Feature

Table 3: Classification results. Evaluation metrics for average Unweighted Accuracy (Accuracy), Sensitivity or True Positive Rate (TPR) and Specificity or True Negative Rate (TNR) are shown. For Human vs Robocalls TPR refers to correct Robocall detection and TNR refers to correct Human detection rate. For Spam vs Non-Spam, TPR refers to correct Spam and TNR to correct Non-Spam detection rate.

	Human vs *Robocall* (%)			*Spam* vs *Non-Spam* (%)		
	TPR	TNR	Accuracy μ ($\pm\sigma$)	TPR	TNR	Accuracy μ ($\pm\sigma$)
OpenSmile L-SVM	91.80	89.34	90.57 ($\pm$2.49)	73.96	67.96	70.96 ($\pm$5.91)
OpenSmile K-SVM	91.05	95.92	93.49 ($\pm$2.12)	70.10	79.57	74.83 ($\pm$5.63)
Log-mel spectrum CNN			-	90.00	75.50	82.75 ($\pm$4.41)

Selection, using both unweighted accuracy and feature weights as a metric for feature selection. We report findings using one of the feature selection methods, Forward Sequential Feature Selection (FW-SFS), since all methods yielded very similar results. We used Python's sci-kit learn *SequentialFeatureSelector* package for this purpose. The L-SVM configuration was used to select best features using the feature weights in the primal problem as a selection metric. We then report performance using K-SVM, averaged over 100 M.C. runs in Fig.1 (red square-marker line). We can see that by using a single feature alone, we can distinguish a *Human* call from a *Robocall* with 79% accuracy, up to 88% accuracy when using a subset of five features. Notice how the standard deviation decreases along the x-axis, as we add more features to the classifier.

The top-scored low-level features selected by FW-SFS are listed here: the mean spectral slope within the 0-500 Hz range for the voiced segments of the recording, with higher values for *Human* calls than for *Robocall*s; the standard deviation of the falling and rising slope for F_0 (pitch), also higher for *Human* calls; pitch jitter variation, higher for *Human* calls; the variation in unvoiced segment length, lower for *Human* calls; and the perceived loudness (50th percentile), with lower values for *Human* calls. Notice that each of the top-5 features provide different attributes of the audio content, including a frequency-based attribute (pitch), spectral attributes (spectral slopes), a temporal attribute (number of unvoiced segments), an energy attribute (loudness). Readers are referred to here,[6] for more details on the calculation of these features.

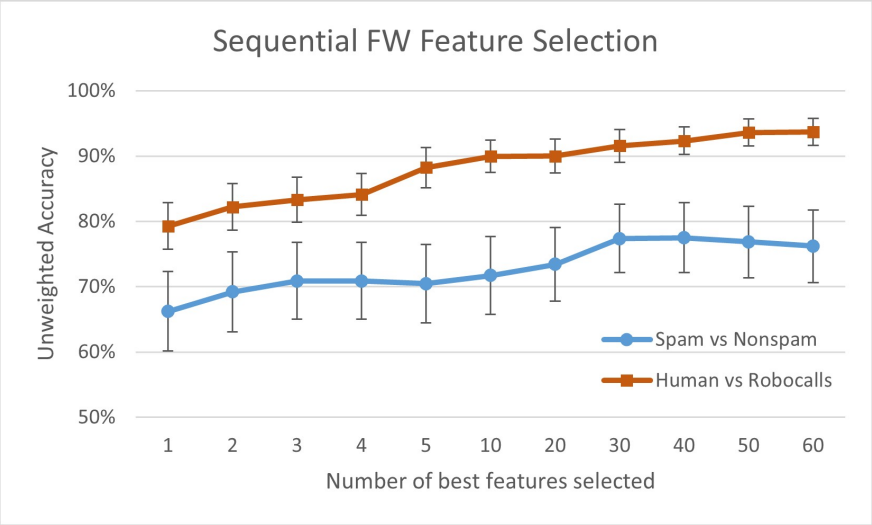

Figure 1: Feature Selection and its impact on classification of Human calls vs Robocalls (red line with square markers) and Spam vs Non-Spam calls (blue line with circle markers), for 100 K-SVM runs.

C. FINDINGS ON *HUMAN SPAM* VS *NON-SPAM* CALLS

In this next part, we will try to answer the following question: Can we distinguish *Spam* calls from *Non-Spam* calls for *Human* callers using audio features alone? An analysis similar to the previous section was performed here using the OpenSmile features and a linear (L-) and an RBF (K-) SVM classifier with fixed parameters $\gamma = 1$ and $C = 1$; once again, we report unweighted (balanced) accuracy over 100 Monte Carlo (MC) runs. Our initial findings in Table 3, columns 4-6 show that distinguishing *Spam* from *Non-Spam* calls is more challenging than distinguishing *Human* calls from *Robocall*s. Using a K-SVM configuration, we achieve 75% accuracy for distinguishing *Spam* from *Non-Spam* calls; this is an 18% drop compared to the task of distinguishing *Human* calls from *Robocall*s at 93%. We additionally invoke a 2D CNN architecture using the log-mel spectrogram as raw features, as described in Section 2.2. As a reminder, for the CNN we only retain 10 sec of the call. The last row of Table 3 shows the average unweighted classification accuracy over 20 random network initializations.

The blue line with circle markers in Fig.1 shows the Sequential FW Feature Selection process for the classification problem of *Human Spam* vs *Non-Spam*. The classification performance increases with the number of added features while its standard deviation remains almost unaffected (5-6%). This analysis further confirms the complexity of this problem compared to classifying *Robocall* from *Human* calls, and the need for an increased number of - or more complex - features. A look into which are the most important features for distinguishing *Human Spam* calls from *Non-Spam* calls reveals a similar trend, as for the *Robocall* vs *Human* calls problem: a similar combination of frequency, spectral, temporal, energy -based features is essentially needed for noticeable increase in performance. By grouping features into types, as shown along the rows of Table 4, we see no clear main contributor. The *feature groups* were created by collecting similar features together, ignoring the imbalance in group sizes: 20 F_0-related features were grouped together, 14 formant-related features were grouped together, 8 spectral slope and spectral ratio features were grouped together etc. The *temporal, rates* group contains temporal or rate feature information including number of voiced segments per sec, average length of voiced and unvoiced segments, number of perceived loudness peaks per sec etc. While the *temporal, rates* and the *perceived loudness* groups seem to be performing lower than other groups, it is worth mentioning again that features from both groups appear in the top best features after the FW-SFS selection process shown previously in Fig. 1, for both classification problems. This fact reinforces the observation that the best selected subset of features would be a combination of features from all groups. Refer here[6] for more details on which features are included into the feature groups.

Table 4: Effect of features groups for classifying Human Spam vs Non-Spam calls, as indicated per row. Unweighted (balanced) Accuracy for K-SVM over 100 M.C runs is shown.

Feature group	Segment Type	Feature #	Accuracy μ ($\pm\sigma$) %
F_0 features	voiced segments	20	64.50 ± 6.56
Formants	voiced segments	14	62.79 ± 7.02
Spectral slope, sp. ratios	voiced segments	8	64.88 ± 6.08
Spectral slope, sp. ratios	unvoiced segments	4	63.12 ± 5.72
Perceived loudness	voiced/unvoiced	10	61.25 ± 6.38
Temporal, rates	voiced/unvoiced	6	53.46 ± 5.29

We finally looked into the importance of accumulative audio-content information, varying the length of the processed recording as a metric. We considered a fixed-length segment, after pre-processing with leading and trailing non-voiced activity removed (see Section 2.2), starting at the beginning of the call and up to N secs. This way, we ensured that the analyzed voice calls begin with voiced segments after pre-processing.

We varied the duration of the call, N, from 0.5 sec to the full recordings and classified the extracted call segments with the OpenSmile K-SVM setup. Fig.2 shows the unweighted (balanced) accuracy performance achieved for varying N secs. Notice that the plateau in the performance curve for longer segment duration can hardly be due to the corpus not having enough calls with long duration: 90% of the *Human Spam* and *Non-Spam* calls have a call duration longer than 10 sec, 69% of them are longer than 20 sec, and 48% are longer than 30 sec. Findings suggest that only a few seconds in a call may be enough to differentiate the two classes using an audio-based approach. This could mean that voice features of call greetings can be a driving indicator of the call type or it could also mean that the presence of background noise is a significant cue. Further exploration is needed to better understand what drives these results, and whether specific background noise or artifacts may be uniquely present in *Spam* or *Non-Spam* calls, potentially affecting current findings.

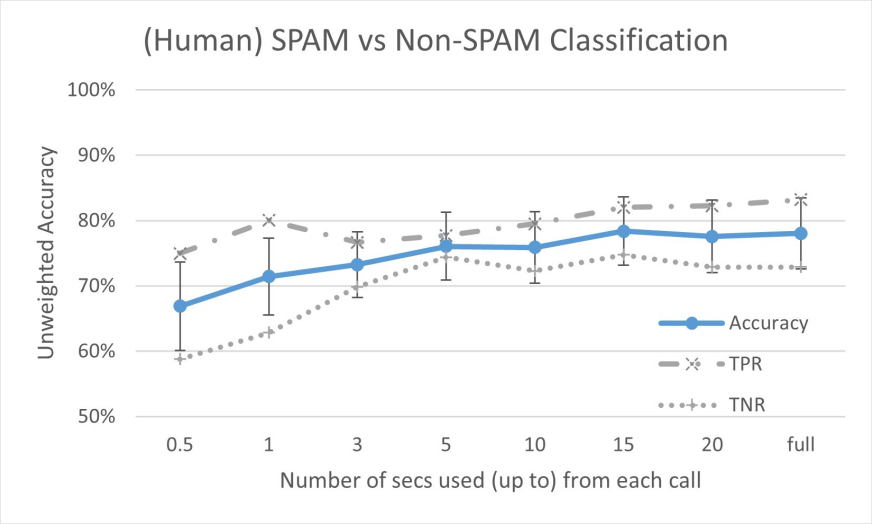

Figure 2: Impact of call length in the classification of Spam vs Non-Spam calls. Unweighted accuracy, True Positive Rate, TPR (Spam), and True Negative Rate, TNR (Non-Spam) are shown, 100 K-SVM runs.

4. CONCLUSION

We proposed an audio-based spam call detection method that uses acoustic features of recorded voicemails to identify *Human* calls from *Robocall*s, and to identify *Spam* calls from *Non-Spam* calls for *Human* callers. To the best of our knowledge, this is the first attempt to explore these tasks with an audio-based methodology that does not rely on the speech content and thus preserves better user-privacy. Results show that voiced and unvoiced audio content carry sufficient discriminatory information to distinguish between *Human* and *Robocall* but also between *Spam* and *Non-Spam* calls. Distinguishing between *Human* calls and *Robocall*s proved to be a less challenging task achieving 93% accuracy, compared to *Spam* vs *Non-Spam* achieving 75-83% accuracy. Looking into the voice gender breakdown of *Human* calls, it appears that a *Spam* call was twice or more likely to come from a (perceived) female voice. We further looked at the speed of annotation, and on average it took 6.6 sec for a reviewer to decide whether a call was *Spam* or *Non-Spam* with reviewers taking slightly longer to label a *Non-Spam* call (7.7 secs) compared to a *Spam* call (6.2 secs). We expect that our automated approach can serve as an auxiliary tool, in combination with other call behavior statistics, to reduce the frequency of unwanted calls or fraudulent incidents. Future work involves augmenting and diversifying the corpus, while exploring acoustic cues from voiced parts of the call.

ACKNOWLEDGMENTS

The authors would like to thank Sebastian Braun from Microsoft Research for providing the state-of-the-art Voice Activity Detector[5] for our work.

REFERENCES

[1] G. Vennila and M. Manikandan, "Detection of human and computer voice spammers using hidden markov model in voice over internet protocol network," *Procedia Computer Science, ICACC*, vol. 115, pp. 588–595, 2017.

[2] M. Amanian, M. H. Yaghmaee Moghaddam, and H. Khosravi Roshkhari, "New method for evaluating anti-spit in voip networks," in *ICCKE 2013*, pp. 374–379, 2013.

[3] A. Natarajan, A. Kannan, V. Belagali, V. N. Pai, R. Shettar, and P. Ghuli, "Spam detection over call transcript using deep learning," in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2* (K. Arai, ed.), (Cham), pp. 138–150, Springer International Publishing, 2022.

[4] C. Pörschmann and H. Knospe, "Analysis of spectral parameters of audio signals for the identification of spam over ip telephony," in *CEAS*, 2008.

[5] S. Braun and I. Tashev, "On training targets for noise-robust voice activity detection," in *European Signal Processing Conference (EUSIPCO)*, August 2021.

[6] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[7] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 2019.