INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Surgical mask detection with deep recurrent phonetic models

Philipp Klumpp[1], *Tomás Arias-Vergara*[1,2], *Juan Camilo Vásquez-Correa*[1,2], *Paula Andrea Pérez-Toro*[1,2], *Florian Hönig*[1], *Elmar Nöth*[1], *Juan Rafael Orozco-Arroyave*[2]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Universidad de Antioquia, Medellín, Colombia

philipp.klumpp@fau.de

Abstract

To solve the task of surgical mask detection from audio recordings in the scope of Interspeech's ComParE challenge, we introduce a phonetic recognizer which is able to differentiate between clear and mask samples.
A deep recurrent phoneme recognition model is first trained on spectrograms from a German corpus to learn the spectral properties of different speech sounds. Under the assumption that each phoneme sounds differently among clear and mask speech, the model is then used to compute frame-wise phonetic labels for the challenge data, including information about the presence of a surgical mask. These labels served to train a second phoneme recognition model which is finally able to differentiate between mask and clear phoneme productions. For a single utterance, we can compute a functional representation and learn a random forest classifier to detect whether a speech sample was produced with or without a mask.
Our method performed better than the baseline methods on both validation and test set. Furthermore, we could show how wearing a mask influences the speech signal. Certain phoneme groups were clearly affected by the obstruction in front of the vocal tract, while others remained almost unaffected.
Index Terms: computational paralinguistics, phoneme recognition

1. Introduction

Against the background of the current global Covid-19 crisis, this year's Interspeech Computational Paralinguistics mask subchallenge [1] provides important results not only to identify whether a person is wearing a surgical mask while speaking. Our main goal is to gain further knowledge how said masks influence the individual speech signal and its intelligibility.
A few related studies have been conducted before, most of them stating that speech intelligibility (SI) strongly depends on the particular type of face mask used. In the following, we term speech of individuals wearing any type of face mask as mask speech. In [2], speech samples had been recorded from participants wearing different variations of healthcare respirators. They found that humanly perceived SI was not significantly affected by surgical masks when compared to a control group. The authors in [3] introduced the Speech Transmission Index (STI) which can serve as a quality measure of SI. As such, it was applied in [4] to show that surgical face masks only have a very limited impact on SI, unlike other types (e.g. filtering facepiece respirators). The study in [5] evaluated SI of speakers wearing surgical masks calculating the perceptual errors of a group of listeners. These listeners were either given an audio-visual signal (a face mask audio signal along with a clear video signal) or an audio-only signal. Their results clearly indicated that the presence of visual information significantly improved perceived SI compared to the audio-only setup. However, it remained unclear whether this difference could be explained with a general influence of visual information on SI which was not related to the obstructed audio signal. The work in [6] made use of the Audio-Visual Face Cover Corpus [7]. Like in [5], they also conclude that surgical masks have a minor impact on SI in a quiet environment, but in their work a lack of visual information decreased SI for both unobstructed and mask speech only in the presence of noise. This finding seems to be supported by numerous studies in the field of automated speech recognition (ASR) that successfully incorporated visual information to improve speech recognition results in noisy environments [8, 9, 10, 11, 12].
The results of above-mentioned previous works have already shown that the impact of surgical masks on a speech signal are rather small. To solve the problem of discriminating between clear and masked speech samples, we have chosen a phonetic approach. On the one hand, this could help us to better understand and explain how our system is working. On the other hand, we could pre-train all our models with a sizable speech corpus and transfer this phonetic knowledge over to the challenge data set.
In the next section, we will give a brief description of the challenge data set and the German speech corpus used to learn a phonetic model. Afterwards, we will highlight the different architectures and the steps taken to transform a phoneme recognizer into a discriminator for clear and masked speech. We then present our results on the development and test subsets and also show how different phonetic groups had been affected by face masks. Finally, we would like to discuss the advantages of the presented method over the approach presented in the baseline paper [1].

2. Materials and Methods

2.1. The mask sub-challenge data set

A brief description of the challenge data is presented. More details can be found in [1]. The Mask Augsburg Speech Corpus (MASC) comprises approximately 10 hours of recorded speech from 16 female and 16 male participants in the age range of 20 to 41. They performed different speech tasks in a quiet environment, once without a mask and once while wearing a mask. The audio was segmented into chunks of 1 second and split into training (10 895 samples / 3.0 h), development (14 647 / 4.1 h) and test set (11 012 / 3.1 h).

2.2. The Verbmobil corpus

To train our phonetic recognizer, we used a subset of the German Verbmobil corpus [13] containing 27 hours of dialogue speech recordings from 593 speakers (307 female, 286 male). The amount of data was doubled by adding Gaussian noise with different SNR (5 dB, 10 dB or 20 dB) to every signal. For our purpose we downsampled the data to 16 kHz using 16 bits/sample and mono-channel configuration. We distributed

http://dx.doi.org/10.21437/Interspeech.2020-1723

Table 1: *Outline of the phoneme recognition model. Output size depended on the length of the sample (T). #c indicates number of channels. #x# denotes kernel size in temporal (first) and frequency (second) domain. [#, #] denotes the stride in the respective domain.*

Output size	Layer
Tx128, 20	20c 1x3 Conv [1, 2]
	Batch Normalization
	LeakyReLU
Tx64, 40	40c 3x3 Conv [1, 2]
	Batch Normalization
	LeakyReLU
Tx32, 60	60c 3x3 Conv [1, 2]
	Batch Normalization
	LeakyReLU
Tx32, 60	Residual Inception Block ch. reduced: 20
Tx16, 100	Reduction Inception Block ch. reduced: 20
Tx16, 100	Residual Inception Block ch. reduced: 40
Tx8, 180	Reduction Inception Block ch. reduced: 40
Tx8, 180	Residual Inception Block ch. reduced: 60
Tx512	512c 1x8 Conv 'valid padding'
	Batch Normalization
	LeakyReLU
Tx400	BiGRU 200 hidden units
Tx400	BiGRU 200 hidden units
Tx31	31c Temporal Conv, window: 3 Softmax activation

the data randomly into training (21 690 files / 48.7 h), development (1168 / 2.8 h) and test (1170 / 2.8 h) sets.

2.3. Data preprocessing

Every audio file was first normalized to a root mean square level of -20 dB. Afterwards, we computed a dual-channel spectrogram. The power spectrum was calculated using a *Hanning* window of 25 ms with 10 ms hop size and 2048 fast fourier transform (FFT) points. A logarithm to the base 10 was applied, then we filtered the spectrogram two times to get the dual-channel result. Similar multi-channel spectrograms have proven beneficial in other applications as well [14]. The first filter-bank was a triangular Mel-bank [15] with 256 bands. The second filter was the same as the first, only with inverted filter-bank order. This would produce an output with high resolutions for higher frequencies and poor resolution in lower frequencies, which is contradictory to the original motivation of the Mel-scale of modeling human auditory perception. In our experiments, however, this inversion helped to improve the results of the mask phoneme recognition.

2.4. Methodology

In the first step, we trained a phonetic recognizer on the Verbmobil corpus with 31 target phonemes (including silence) and a temporal resolution of 10 ms. The phoneme labels required for this were generated by force-aligning the transliteration using a traditional GMM-HMM recognizer trained on the same material, using Kaldi [16]. In the next step, we used this

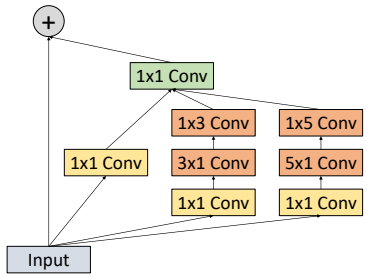


Figure 1: *Schematic of the residual inception block used in the phoneme recognition network.*

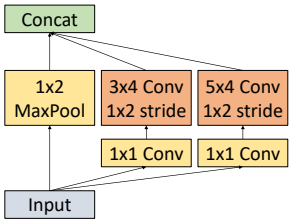


Figure 2: *Schematic of the reduction inception block used in the phoneme recognition network.*

recognizer to compute framewise phonetic labels for the audio files of the mask challenge. For every phoneme, we made the assumption that it could be produced while wearing a mask or not, resulting in two different target phonemes. Hence, our original phoneme space was increased to 61 targets: Silence (assumed not to change between clear and mask samples), 30 phonemes without mask and the same 30 phonemes with mask. For every file, we then had to find the phonemes and map them to either their clear or mask representation. With phonetic labels available, we could train a second phoneme recognizer, this time with 61 phonetic targets and our training data from the mask challenge. Afterwards, we could use this recognizer to perform a phoneme recognition on unseen data which would not only determine phonetic posteriors, but would also provide information whether a phoneme was produced while wearing a mask or not. After computing framewise phonetic posteriors, we computed several functionals which represented the overall phonetic posteriors of a whole utterance and trained a random forest classifier to assess whether an utterance was spoken with or without mask.
When we computed results for the test set, we trained our phoneme recognizer with the training and the development set as it was done for the baseline as well. However, we left out a random but constant selection of 1000 files from the original development set to evaluate training progress.

2.5. Deep recurrent phoneme recognition

The phoneme recognition model was comprised of a convolutional part for feature extraction from the spectrogram and a recurrent part for sequential phoneme classification. The whole network architecture is shown in Table 1. After the first three convolutional layers which served to reduce the number of frequency bands, we applied two types of convolutional blocks that were inspired by the inception model [17]. The core idea was to put multiple convolution kernels of different sizes in parallel to

Table 2: *Outline of the final layers of the mask phoneme recognition model which were appended to the convolutional part of the original phoneme recognizer.*

Output size	Layer
	Dropout 60 %
Tx500	BiGRU 250 hidden units
	Dropout 60 %
Tx500	BiGRU 250 hidden units
	Dropout 60 %
Tx61	61c Temporal Conv, window: 3 Softmax activation

allow for the varying temporal and frequency patterns of different phonemes. Our residual inception block shown in Figure 1 first performed a channel reduction with 1x1 convolutions. Afterwards, two separate filter kernels applied a convolution over the time and frequency domains. This architecture of channel reduction and separate filter kernels is also referred to as depthwise separable convolutions. It was introduced in [18] and has proven effective in architectures such as MobileNet [19] to help reduce the number of parameters and the overall computational complexity of a neural network. The results of the three convolution branches were concatenated along the channel dimension and the input channel configuration was restored with another 1x1 convolution. The result was added to the original input to realize a residual connection [20].
The reduction inception block in Figure 2 was used to collapse the frequency dimension further. It was also inspired by the inception architecture and applied a max pooling along the frequency bands. In parallel to the pooling, it comprised two 1x1 convolutions for channel reduction and afterwards two strided convolutions to perform a convolutional downsampling. The results of all three layers were finally concatenated along the channel dimension. We used leaky rectified linear unit (ReLU) activation ($m = 0.3$) after every residual or reduction inception block.
After the convolutional part, we added a stack of two bidirectional recurrent layers using Gated Recurrent Unit [21] (GRU) cells with 200 hidden units. Both the forward and the backward pass were configured to return sequential output and their results were concatenated for every time step. The final layer of the network was a temporal convolution with window size three and 31 output channels, one for each phonetic class. We applied softmax activation to the outputs of the final layer to get posterior probabilities.
The network was trained using Adam optimizer [22] with an initial learning rate of 0.001 ($\beta_1 = 0.9, \beta_2 = 0.999$), a global learning rate decay of 0.95 per epoch, and categorical entropy loss. We applied $L2$ kernel regularization to every convolution except the final classification layer. We considered an early stopping strategy with a patience of three epochs.

2.6. The mask phoneme recognition model

After we used the recognizer explained in section 2.5 to compute framewise phonetic labels for the challenge data set, we modified the phoneme recognizer such that it could predict the new 61 phonemes (1 silence, 30 mask phonemes and their 30 clear phonetic counterparts). We used the pre-trained model to have a decent initialization for the convolutional layers, removed the recurrent layers as well as the final classification layer and replaced them with a new stack of layers shown in Table 2. We increased the number of hidden units in the recurrent layers slightly to 250 due to the higher complexity of the new classification task. Additionally, we applied a dropout of 60 % to the output of the convolutional part and every recurrent layer. We did this to prevent overfitting because the mask data set was considerably smaller compared to the Verbmobil corpus.
The training setup was identical to the one presented in 2.5. We added class weights to the loss function to lower the importance of the silence phoneme which received a factor of 0.2, the other phonemes remained unchanged with weights of 1. We motivated this step by the fact that silence was overrepresented in the challenge data compared to the other phonemes. This is not surprising for the task of phoneme recognition, but putting less emphasis on the correct classification of silent segments would help the network to better learn to classify the other phonemes.

2.7. Final feature vector and classification

We used the trained mask phoneme recognition model to perform frame-wise classification of the computed spectrograms. For every of the 98 frames per sample, we got 61 posteriors predicting which phoneme this frame contained and whether it was spoken with or without a mask (in case it was not a silent segment). To create a single representation for every sample, we decided to compute the following functionals: The mean posterior probability of every phoneme, the maximum posterior probability, the mean difference for every phoneme between its mask and clear posterior values (a positive result would indicate a clear phoneme and vice versa, not computed for silence, resulting in 30 values) as well as the maximum sequence length of consecutive mask or clear phonemes. The final feature vector contained 154 values.
To classify every sample, we trained a random forest classifier with 100 decision trees, maximizing the information gain for every split and applying class weights to balance the training data distribution. We used this final classifier to predict whether an unseen sample was produced with a mask or a clear mouth.

2.8. Phonetic analysis of clear and mask speech

To evaluate how certain phoneme groups were affected by wearing a mask, we extracted the LeakyReLU activation after the convolutional layers which served as input to the first recurrent layer. We chose this particular activation because it comprised the feature extraction from the convolutional neural network (CNN), without inducing future or past context through the recurrent neural network (RNN). Activations were collected for all frames that had the same phoneme label in the two neighboring frames, such that phoneme boundaries were not covered. Every 512-dimensional activation vector was assigned to one of eight phoneme groups: Open or closed vowels, fricatives, nasals, voiced or unvoiced plosives, approximants and vibrants. We performed principal component analysis (PCA) to reduce the activation vector's size to one. Finally, the difference in mean and variance between clear and mask PCA results were computed to determine if certain phoneme groups were more affected by wearing a mask than others.

3. Results

3.1. Phoneme recognition results

To provide a better insight of the difficulty of the phoneme recognition task for clear and mask speech, we provide a brief summary of the two models. The first one, trained on the Verbmobil corpus, achieved an overall accuracy of 81.4 % (31 classes) per frame. The accuracy of the second model which was trained to differentiate between clear and mask phonemes

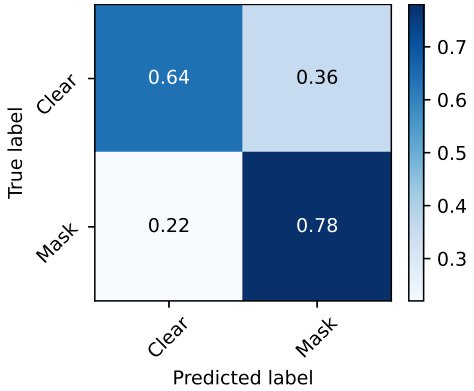
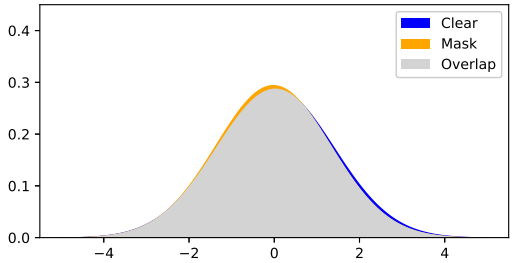


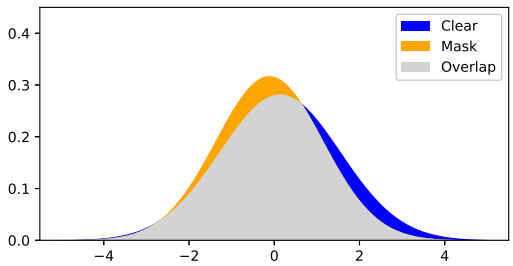
Figure 3: *Confusion matrix of classification results achieved on the development set.*

Table 3: *Differences between clear and mask speech in mean and variance for PCA results of all phoneme groups.*

Phonetic group	$\Delta\mu$	$\Delta\sigma^2$
Open vowels	0.00	0.13
Closed vowels	0.07	0.04
Fricatives	0.21	0.26
Nasals	0.01	0.23
Voiced plosives	0.05	0.09
Unvoiced plosives	**0.25**	**0.42**
Approximants	0.21	0.01
Vibrants	0.18	0.01



(a) *Voiced plosives*



(b) *Unvoiced plosives*

Figure 4: *Distribution of PCA results for convolutional activations in the phoneme recognition model for voiced and unvoiced plosives. The unvoiced plosives show a more significant difference between the distribution of clear and mask speech.*

was significantly lower with 54.9 % (61 classes).

3.2. Classification results

Our architecture achieved an unweighted average recall (UAR) of 70.8 % on the development set, which was significantly better than the best reported result of the baseline method (64.4 %). For better visualization of the class-specific performance, we have provided a confusion matrix in Figure 3. After we included the majority of the development data in the training and evaluated the test set, we achieved 75.4 % UAR (baseline 71.8 %).

3.3. Phonetic results

To better understand how wearing a mask affected certain phonemes, we analyzed the intermediate activations after the convolutional part of our model. After evaluation of the PCA results for different phoneme groups (Table 3), we found that certain categories were almost not affected. Such clusters were voiced plosives as well as both open and closed vowels. The groups that showed the most pronounced deviation between clear and mask speech in mean and variance of PCA results were unvoiced plosives and fricatives. Both differences were found to be clearly significant through a t-test (p-values $\ll$ 0.001). In Figure 4 we plotted the normal distributions of results for voiced (/b/, /d/, /g/) and unvoiced plosives (/p/, /t/, /k/). For voiced plosives, there is no significant difference in the distributions of clear and mask samples. However, we can see such difference in the plot of their unvoiced counterparts.

4. Discussion

The major decrease in frame-wise classification accuracy for the mask phoneme recognition model showed the increased complexity of the new task. Instead of only classifying phonemes, the model now had to differentiate between phoneme productions with and without a mask.
When we used this model to determine whether a whole sample was spoken with a mask or not, we could show that this differentiation was well above chance level (50 %). Our method performed well on both the provided development and test sets. With a higher degree of success than all baseline methods, we were able to distinguish between clear and mask speech. We could also support our results by showing how different phoneme groups were affected by an obstruction in front of a speaker's mouth. The observed difference is not large, but this conforms to the findings from previous works [2, 4, 5, 6]. Nevertheless, for certain phonemes like fricatives and unvoiced plosives the difference is big enough to help differentiate between clear and mask speech. This seems plausible, because both groups require an unobstructed flow of air out of the vocal tract. For fricatives, this is a constant flow of air (e.g. in /f/, /s/ or /ʃ/), whereas for the unvoiced plosives, it is a sudden, powerful burst after a closure (e.g. in /p/, /t/ and /k/).

5. Conclusion

With a phonetic representation of clear and mask speech, we were able to outperform well-established acoustic feature extraction tools [23] as well as deep feature and representation learning approaches presented in the baseline paper [1]. Additionally, our method provided further insights into the morphology of mask speech in comparison to clear speech, showing how different groups of phonemes were affected by a surgical mask.

6. References

[1] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, September 2020, p. 5 pages, to appear.

[2] L. J. Radonovich Jr, R. Yanke, J. Cheng, and B. Bender, "Diminished speech intelligibility associated with certain types of respirators worn by healthcare workers," *Journal of Occupational and Environmental Hygiene*, vol. 7, no. 1, pp. 63–70, 2009.

[3] T. Houtgast and H. J. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acta Acustica united with Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[4] A. J. Palmiero, D. Symons, J. W. Morgan III, and R. E. Shaffer, "Speech intelligibility assessment of protective facemasks and air-purifying respirators," *Journal of occupational and environmental hygiene*, vol. 13, no. 12, pp. 960–968, 2016.

[5] C. Llamas, P. Harrison, D. Donnelly, and D. Watt, "Effects of different types of face coverings on speech acoustics and intelligibility," 2009.

[6] N. Fecher and D. Watt, "Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[7] N. Fecher, "The" audio-visual face cover corpus": Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[8] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[9] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7596–7599.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[11] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[12] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[13] W. Wahlster, *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

[14] T. Arias-Vergara, J. C. Vasquez-Correa, S. Gollwitzer, J. R. Orozco-Arroyave, M. Schuster, and E. Nöth, "Multi-channel convolutional neural networks for automatic detection of speech deficits in cochlear implant users," in *Iberoamerican Congress on Pattern Recognition (CIARP)*. Springer, 2019, pp. 679–687.

[15] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[18] L. Sifre and S. Mallat, "Rigid-motion scattering for image classification," *Ph. D. thesis*, 2014.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

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

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

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

[23] 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*, 2010, pp. 1459–1462.