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

Phonetic, Frame Clustering and Intelligibility Analyses for the INTERSPEECH 2020 ComParE Challenge

Claude Montacié[1] *and Marie-José Caraty*[2]

[1]STIH Laboratory, Sorbonne University, 28 rue Serpente Paris, 75006, Paris, France
[2]STIH Laboratory, Paris University, 45 rue des Saints-Pères, 75006, Paris, France
claude.montacie@sorbonne-universite.fr, marie-jose.caraty@parisdescartes.fr

Abstract

The INTERSPEECH 2020 Compare Mask Sub-Challenge is to determine whether a speech signal was emitted with or without wearing a surgical mask. For this purpose, we have investigated phonetic context and intelligibility measurements related to speech changes caused by wearing a mask. Experiments were conducted on the Mask Augsburg Speech Corpus (MASC) and on the Mask Sorbonne Speech Corpus (MSSC) both in German language. We investigated the effects of mask wearing on the acoustical properties of phonemes at frame and segment levels. At the frame level, a phonetic mask detector has been developed to determine the most sensitive phonemes when wearing a mask. At the segmental level, a perceptual scoring of intelligibility has been developed and assessed on the MSCC. Two mask detector systems have been developed and assessed on the MASC: the first one used two large composite audio feature sets, the second one used a bottom-up approach based on phonetic analysis and frame clustering. Experiments have shown an improvement of 5.9% (absolute) on the Test set compared to the official baseline performance of the Challenge (71.8%).
Index terms: Computational Paralinguistics, Challenge, Face covering, Surgeon mask, Intelligibility measurement

1. Introduction

The sentinel Lite mask is the standard chirurgical face mask used for the Mask Sub-Challenge (MASC) [1]. The medical mask was designed to avoid the projection of droplets of the mask wearer in the air. Secretions from saliva or the upper airways, droplets may contain infectious agents as air. So wearing a mask protects the patient and the environment from contamination.

Forensic Sciences have a great interest in masked speaker identification [2] or recognition [3]. Indeed, many crimes or misdemeanors were committed by individuals wearing a mask. There is a wide range of masks that can be used for such a purpose. Regarding the impact on communication and acoustic consequences of wearing a mask, there are two important characteristics of a mask: the hidden face of the speaker and the material it is made of. Eight types of face coverings: motorcycle helmet, balaclavas, niqāb, surgical mask, hoodie/scarf and rubber mask are illustrated in [2] showing the diversity of the masks. A mask is typically made of fabric or polymer [4]. Depending on these characteristics, a greater or lesser degradation of speech perception and intelligibility was observed [5, 6].

Auditory and visual information interact during speech communication [7]. Wearing a face mask (without openings at mouth and nose) affects communication between the speaker (mask wearer) and the listener in two ways for the speaker and the listener. For the speaker, obstruction of the mouth and nose by the mask with possible articulatory discomfort in the lips and jaw leads to an unusual audio feedback of his or her speech. In the concept of the audio-phonation loop [8], this feedback induces on the one hand a reflex adaptation of the mask wearer consisting of an increase in vocal intensity and on the other hand more intentionally a better articulation or possibly a hyper-articulation to preserve intelligibility for the listener. Similarities in speaker adaptation can be found with the Lombard effect [9] when a speaker is in a noisy environment and with the addressing effect [10] when a speaker addresses a child or a foreigner with selected prosody and speech articulation. For the listener, the lack of visual cues such as visibility of the mouth and facial expressions of the mask wearer leads to a degradation of perception and even intelligibility [7, 11, 12].

Much of the related work on the wearing a mask when speaking focuses on changes in the acoustics of speech signals and their intelligibility [11, 12, 13, 14, 15]. Numerous perceptive experiments aimed at analyzing the effect of the mask wearing from listening errors on dedicated databases such as the audio and audio-visual face cover corpus [14]. The effect of masks on the spectral properties of phonemes was studied in [2, 12, 15]. For example, three main confusions were found: the stops /t/ and /f/ with /θ/, and the nasal /n/ with /ŋ/ [2]. Changes in the acoustics of speech signals were measured using spectral transfer function [3, 11] or objective scoring of the intelligibility [16]. For the material a mask is made of and as for fabrics [4], it is possible to study the flow resistance, the sound absorption coefficient and its effects on the spectrum of speech transmitted through material [17].

For the INTERSPEECH 2020 ComParE Mask Sub-Challenge and on the basis of related work, we paid particular attention to phonetic context and its influence in the detection of speech changes caused by wearing a mask. Thus, we have developed a bottom-up detection approach based on phonetic analysis and frame clustering. The document is organized as follows: Two speech databases are presented in Section 2: Mask Augsburg Speech Corpus provided for the Challenge [1] and Mask Sorbonne Speech Corpus. Our Mask baseline system is described in Section 3 and was used as a reference system in the next Sections. In Section 4, a frame level mask phonetic detector is presented and assessed on the Devel set using a bottom-up approach. This approach is also used with a frame level mask cluster detector in Section 5. Speech intelligibility measurement using an acoustic phonetic decoding system is described in Section 6. Experiments on the Test set are presented in Section 7. The last Section concludes the study.

http://dx.doi.org/10.21437/Interspeech.2020-2243

2. Speech databases

Two speech databases were used for the experiments. The first one is the Mask Augsburg Speech Corpus (MASC) provided for the Challenge [1] and the second one is the Mask Sorbonne Speech Corpus (MSSC) recorded for measuring speech changes caused by sound absorption related to the mask material.

2.1. Mask *Augsburg* Speech Corpus

The Mask Augsburg Speech Corpus (MASC) consists of 36,554 non-overlapping one-second speech chunks. This corpus was divided into Train, Devel and Test sets. For the whole database, 32 German were recorded with or without chirurgical mask wearing. No textual transcript is provided and it appears that no two speech chunks have the same vocal content. 3,509,185 frames (96 frames per speech chunk) were obtained using short term centisecond analysis on MASC. Table 1 shows the occurrence of phonemes and diphthongs as a percentage of the 1,045,021 frames of the MASC Train set. International Phonetic Alphabet was used for the transcription of the phonemes. The transcription of the corpus was obtained by an unconstrained acoustic-phonetic decoding [18].

Table 1: *Percentage of the number of frames (%) per Training set phonetic class.*

Front vowels (#113,799)							
ɪ	i	ɛ	e	y	œ	ɤ	ø
2.3	1.5	1.4	1.3	0.5	0.5	0.4	0.3
Central vowels (#182,786)							
ə	ɐ	a	ɑ				
4.9	4.2	3.1	2.5				
Back vowels (#46,444)							
o	ʊ	ɔ	u				
1.3	1.3	1.1	0.8				
Diphthongs (#32,426)							
aɪ	aʊ	ɔɤ					
2.0	0.8	0.3					
Voiced fricatives (#61,775)							
z	ʁ	v	j	ʒ			
2.2	1.9	1.3	0.5	0.02			
Unvoiced fricatives (#124,281)							
s	f	ʃ	ç	h	x		
3.7	3.1	1.8	1.7	1.1	0.5		
Voiced plosives (#46,755)							
d	b	g	dʒ				
2.1	1.2	1.1	0.1				
Unvoiced plosives (#130,521)							
t	ʔ	ts	k	p	pf	tʃ	
4.6	2.3	2.0	1.6	1.1	0.7	0.2	
Nasals (#129,065)							
n	m	ŋ					
7.4	4.1	0.8					
Lateral (#32,695)							
l							
3.1							

The phonetic classes have been grouped into 10 macroclasses: front, central and back vowels (3); diphthongs (1), voiced and unvoiced fricatives (2), voiced and unvoiced plosives (2), nasals (1), lateral (1). All selected macroclasses have sufficient frame occurrences in the Train set to learn, if possible, a model to distinguish between masked and unmasked frames.

2.2. Mask *Sorbonne* Speech Corpus (MSSC)

The Mask Sorbonne Speech Corpus (MSSC) consists in 2,000 recorded sentences in German language. In order to measure speech changes caused by sound absorption by the mask material, we used an artificial voice generator held in a block of high-density foam. The voice generator was a micro bluetooth speaker (Bose Sound Link micro) located in an anechoid chamber. 1,000 sentences uttered by 30 speakers have been selected from a speech data corpus: the German Distant Speech Data Corpus [19]. Kinect Direct Access recording conditions have been chosen. The list of selected sentences is available on [20]. The speech signal was transmitted to the voice generator and was recorded using a digital sound recorder (Zoom H4n Pro) at a rate of 48 kHz with 24 bit, downsampled and converted to 16 kHz and mono/16 bit. Two recording conditions were used: a voice generator without mask and a voice generator with surgical mask. There were two recording sessions: one for the Train set (16 speakers, 1,068 sentences) and one for the Devel set (14 speakers, 932 sentences).

3. Mask Baseline system

For the development of the Mask detector, we have chosen for the Mask Basic System (MBS) a linear Kernel SVM classifier with a composite audio feature set (8,016 features) called MBS set. This audio set is the concatenation of the auDeep-fused feature set (4,096 features), the Bag-of-Audio-Words feature set (2,000 features) and the DeepSpectrum feature set (1,920 features) using as pretrained CNN DenseNet201. Four Toolboxes [1] were used to compute audio features: auDeep, OpenSmile [21], OpenXbox and DeepSpectrum. These usual audio feature sets allow a representation of audio files in terms of spectral, cepstral, prosodic and voice quality information. Scikit-learn toolbox [22] was used for the implementation of the SVM classifier. Posterior probabilities were computed by the isotonic regression method [23]. Complexity parameters of the SVM classifier were optimized to maximize the Unweighted Average Recall (UAR) on the Devel set. The Mask detector gave an UAR of 67.4% on the Devel set of MASC. The UAR on the Devel set of MSSC corpus is 88.2%. For both experiments, the mask detector was trained on their respective Train set. A merging of the two Train sets was tested and gave an UAR of 68.4% on the Devel set of MASC compared to the higher UAR on the Devel set 64.4% referred in [1]. The very good result on the MSSC corpus obtained by the voice generator seems to show that the effects due to sound absorption by the mask seem to be more important than those due to an audio-phonation loop. Indeed, speech changes on MSSC are related only to sound absorption by the mask while on MASC they are related to sound absorption and audio loop.

4. Bottom-up detection approach and phonetic analysis

We have chosen to use a bottom-up approach for mask detection. It consists in building a mask detector at the frame level taking into account the phonetic context and using the set of decisions obtained for each frame to elaborate a decision at the chunk level. We chose to explore this approach for two

reasons: 1) previous studies showing that the effects of a mask on speech depend on the phonetic context [2], 2) the sound absorption of the mask and its effect on the speech spectrum occurs at the frame level [17].

4.1. Frame level mask generic system

For the development of the frame level generic Mask detector, we have a linear Kernel SVM classifier with the 130 Low Level Descriptors of ComParE [24]. The frame level Mask detector gives an UAR of 54.7% on the frames of the Devel set of MASC. The decision at the chunk level is obtained from the posterior average computed at the frame level. The chunk level Mask detector gave an UAR of 59.6% on the Devel set of MASC.

4.2. Frame level mask phonetic system

For the development of the frame level phonetic Mask detector, we have eleven linear Kernel phonetic-based SVM classifiers, one for each of the ten phonetic macroclasses plus one for the silence macroclass. Each of the phonetic-based SVM classifier was trained on the frames whose phonetic label belongs to the phonetic macroclass. The 130 Low Level Descriptors of ComParE have been used. The frame level phonetic Mask detector gave an UAR result of 56.8% across the MASC Devel set, a 2% improvement over the generic system. The decision at the chunk level was obtained from the posterior average computed at the frame level. The chunk level Mask detector gave an UAR of 63.9% on the Devel set of MASC: a 4.5% improvement over the generic system.

Table 5 shows the UAR results on the frames of the Devel set by phonetic macroclasses and phonemes. For each macroclass, the first line gives the number of frames belonging to this macroclass in the Devel set as well as the UAR obtained on these frames by the phonetic models, the second line gives the list of phonemes of the macroclass, the third line (in italics) gives per phoneme the UAR obtained by the generic model, the fourth line (in bold) gives per phoneme the UAR obtained by the phonetic model.

It can be noticed that in the majority of cases, phonetic models allow a better mask wearing detection than the generic model. Only the frames belonging to the phonetic classes /ʒ/ /g/ /ʔ/ /tʃ/ /ŋ/ allow a less good mask wearing detection by the phonetic model than by the generic model. However, these frames represent only 4.5% of all the frames. We also notice that the frames belonging to the macro classes Diphthongs, Laterals, Central and Back vowels allow an above-average mask detection (56.8%). It could be an indicator that phonemes belonging to these macroclasses are more sensitive to the mask wearing.

5. Frame Clustering analysis

Chunk level mask Phonetic system has led to a 4.5% improvement in performance. We chose to build N homogeneous clusters of frames using the k-means algorithm [25] in order to improve the performance of the mask detector. For the development of the frame level clustering Mask detector, we have used N linear Kernel cluster-based SVM classifiers, one for each cluster. Each of the cluster-based SVM classifier was trained on the frames belonging to the cluster. To take into account temporal evolution, a couple of 130 Low Level Descriptors of ComParE [24] computed from two consecutive short-term analyses have been used to represent a frame. Table 2 gives the UAR on the Devel set of MASC at frame and chunk levels for a number of clusters ranging from 16 to 1024.

Table 2: *Frame and Chunk UARs on the Devel set based on the cluster models*

Number of Clusters	Frame UAR (%)	Chunk UAR (%)
16	57.5	64.4
32	57.8	64.9
64	58.1	65.9
128	58.1	67.0
256	58.1	67.5
512	57.9	68.0
1024	57.6	68.2

We notice that the frame-level UAR has a maximum for 256 clusters but that the chunk-level UAR continues to improve up to 1024 clusters. The chunk level Mask detector gives an UAR of 68.2 % on the Devel set of MASC, a 4.3% improvement over the phonetic system. Fusion with the Mask baseline system using average of posteriors or posterior probabilities gave an UAR of 70.4%.

6. Speech Intelligibility measurement

Measures of speech intelligibility [26, 27] are used in many areas of speech processing such as speech coding and synthesis speech reinforcement and enhancement. These measures are also used to assess the influence of mask wearing on speech. Speech intelligibility can be measured by perceptual methods such as listening tests [28] or by objective methods [29] based on signal analysis. Perceptual approach has generally been considered to be more reliable for complex environmental conditions. We used both types of approaches to study speech changes on the MSSC corpus.

6.1. Perceptual scoring of the intelligibility

Many methods have been standardized to measure intelligibility by auditors [30]. These include Diagnostic Rhyme Test (DRT) with rhyming monosyllable word pairs listening and Modified Rhyme Test (MRT) with rhyming monosyllable word sets listening. We have chosen to develop an automatic method based on the comparison of the phonetic transcriptions of two recordings with the same vocal content: the first one from the voice generator without mask and the second one from the voice generator with surgical mask. The alignment of the two transcripts was done by the Wagner and Fisher algorithm [31]. This method was used on all 1,000 pairs of sentences in the MSSC corpus. It yielded 69,576 pairs of phonemes or diphthongs, 24,688 of which corresponded to confusions, i.e. a confusion rate of 35%. 16,571 confusions occur between phonemes of different macroclasses and 8,117 between phonemes belonging to the same macroclass. The confusions between phonemes of different macroclasses are in order according to the rate of confusion: y/ə (13%), ʏ/ə (10%), ɔʏ/ɐ (9%), ʏ/l (8%), ø/ə (8%), v/b (7%), f/t (6%), ʊ/ə (6%) and ɪ/ə (5%). There is a lot of confusion between front vowels (y, ʏ, ø, ɪ) and the central vowel /ə/ as well as between the fricatives (/v/, /f/) and plosives (/b/, /t/). There are more unexpected confusions such as confusion between the front

vowel /ʏ/ and the lateral /l/. Table 3 gives the confusion rate between the phonemes of a macroclass for each macroclasse.

Table 3: *Confusion rate inside a macroclass*

Macroclasses	Average intra-macroclass confusion rate
Front vowels	**20.4%**
Central vowels	15.0%
Back vowels	**21.5%**
Diphthongs	1.2%
Voiced fricatives	6.6%
Unvoiced fricatives	12.7%
Voiced plosives	10.1%
Unvoiced plosives	17,4%
Nasals	**20.3%**

We notice that confusion rate between the phonemes of a macroclass is the highest for the following macroclasses: backs vowels, front vowels and nasals.

6.2. Objective scoring of the intelligibility

Many methods have been standardized to measure intelligibility by signal processing [32]. These methods include the Speech Intelligibility Index (SII) measure [32] based on the estimation of the weighted average of the signal-to-noise ratios (SNR) in several spectral bands and the Speech Transmission Index (STI) measure [32] based on modeling of convolutional and additive noise from the reduction in temporal envelope modulations. All these measures range from 0 (complete loss of intelligibility) to 1 (no change in intelligibility). We have chosen to use another measure of intelligibility, the Short Term Objective Intelligibility measures (STOI), for which a toolbox allowing its computation is available [33]. The average intelligibility measurement between MSSC recordings without mask and MSSC recordings with surgical mask was 0.95. This confirms previous studies that show that wearing a surgical mask only slightly reduces speech intelligibility [11].

7. Experiments on the Test set

The official baseline of the Mask Challenge on the Test set is 71.8% in terms of UAR. We chose for the experiments a Mask detector based on the fusion of two systems. The first system uses two large composite audio feature sets: MBS set (8,016 features) and MBS set concatenated to Resnet50 (10,064 features). The second system uses a bottom-up approach based on frame clustering (256 and 1024 clusters) and phonetic analysis. Table 4 gives the UAR results of the Mask detector.

Table 4: *UAR results (%) on the Test set*

#sub	audio feature set	bottom-up approach	UAR
1	MBS set	Phon. + 1024 clusters	76.3%
2	MBS set	1024 clusters	76.2%
3	MBS set	Phon. + 256 clusters	76.9%
4	MBS set + resnet50	Phon. + 256 clusters	**77.7%**

The best submission in terms of UAR was 77.7% on the Test set. This is a significant improvement of 5.9% absolute.

8. Conclusion

In this paper, phonetic context related to speech changes caused by wearing a mask has been investigated. A speech database based on an artificial voice generator has been developed to measure speech changes caused by sound absorption by the mask material. A perceptual scoring of intelligibility based on comparison of the acoustic phonetic decoding has been developed and assessed on this database. The phonemes the most sensitive to the mask wearing were the nasals, the front and back vowels. Two mask detector systems have been developed and assessed: the first one using a large composite audio feature set and the second one using a bottom-up approach based on phonetic analysis and frame clustering. Experiments with a fusion of the two detectors have shown an improvement of 5.9% on the Test set compared to the official baseline performance of the Challenge (71.8%).

Table 5: *UAR results (%) on the frames of Devel set by phonetic macroclasses and phonemes.*

Front vowels (#160,040)							**UAR. 56.8%**
ɪ	i	ɛ	e	y	œ	ʏ	ø
54.4	*53.7*	*54.4*	*56.2*	*50.2*	*55.2*	*55.4*	*51.1*
57.0	**55.2**	**56.9**	**58.9**	**51.5**	**58.1**	**55.8**	**53.0**
Central vowels (#211,135)							**UAR. 57.8%**
ə	ɐ	a	ɑː				
55.1	*55.0*	*55.9*	*56.7*				
56.6	**57.8**	**58.2**	**59.7**				
Back vowels (#59,499)							**UAR. 57.8%**
o	ʊ	ɔ	u				
57.7	*55.7*	*57.8*	*54.6*				
58.6	**58.0**	**58.1**	**54.9**				
Diphthongs (#52,463)							**UAR. 59.2%**
aɪ	aʊ	ɔʏ					
55.3	*58.0*	*55.4*					
59.3	**60.8**	**56.1**					
Voiced fricatives (#91,795)							**UAR. 55.4%**
z	ʁ	v	j	ʒ			
55.4	*54.8*	*54.9*	*54.2*	*55.4*			
55.9	**55.2**	**56.0**	**54.8**	**51.3**			
Unvoiced fricatives (#165,953)							**UAR. 56.8%**
s	f	ʃ	ç	h	x		
55.8	*55.0*	*53.0*	*53.6*	*55.1*	*54.6*		
57.6	**55.6**	**55.2**	**55.8**	**55.1**	**56.2**		
Voiced plosives (#60,395)							**UAR. 54.8%**
d	b	g	dʒ				
54.3	*54.7*	*55.2*	*53.3*				
54.7	**55.1**	**54.4**	**54.8**				
Unvoiced plosives (#180,958)							**UAR. 56.3%**
t	ʔ	ts	k	p	pf	tʃ	
55.0	*54.8*	*57.2*	*55.2*	*56.2*	*55.4*	*56.2*	
56.3	**54.3**	**57.8**	**55.7**	**57.3**	**55.9**	**54.8**	
Nasals (#181,913)							**UAR. 55.7%**
n	m	ŋ					
55.6	54.4	59.8					
55.8	**55.0**	**56.7**					
Laterals (#42,733)							**UAR. 58.3%**
l							
54.1							
58.3							
Total (#1,406,112)							**UAR. 56.8%**

9. 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. Deighton MacIntyre and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks", Proceedings INTERSPEECH 2020, ISCA, Shanghai, China, 2020.

[2] N. Fecher, Doctor of Philosophy Thesis, University of York, "Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants", 414 pages, 2014.

[3] R. Saeidi, I. Huhtakallio and P. Alku, "Analysis of Face Mask Effect on Speaker Recognition", INTERSPEECH 2016, ISCA, San Francisco, USA, pp. 1800–1804, 2016.

[4] K. P. Chellamani, D. Veerasubramanian and R. V. Balaji, "Surgical face masks: manufacturing methods and classification", Journal of Academia and Industrial Research, 2, pp. 320–324, 2013.

[5] L. L. Mendel, J. A. Gardino and D. S. R. Atcherson, "Speech Understanding Using Surgical Masks: A Problem in Health Care?", Journal of the American Academy of Audiology, vol. 1, no. 3, pp. 686–695, 2008.

[6] K. J. Wittum, L. Feth and E. Hoglund "The effects of surgical masks on speech perception in noise", Proc. Mtgs. Acoust. 19, 060125; doi: 10.1121/1.4800719, Acoustical Society of America, 2013.

[7] T. Kawase, K. Yamaguchi, T. Ogawa, K. I. Suzuki, M. Suzuki, M. Itoh and T. Fujii, "Recruitment of fusiform face area associated with listening to degraded speech sounds in auditory–visual speech perception: A PET study", Neuroscience letters, vol. 382, no. 3, pp. 254–258, 2005.

[8] M. Garnier, N. Henrich and D. Dubois, "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect", Journal of Speech, Language, and Hearing Research, vol. 53, pp. 588–608, 2010.

[9] R. Marxer, J. Barker, N. Alghamdi and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems", Speech communication, vol. 100, pp. 58–68, 2018.

[10] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants", Developmental Review, vol. 27, no. 4, pp. 501–532, 2007.

[11] C. Llamas, P. Harrison, D. Donnelly and D. Watt, "Effects of different types of face coverings on speech acoustics and intelligibility", pp. 80–104, 2009.

[12] N. Fecher and D. Watt, "Speaking under cover the effect of face-concealing garments on spectral properties of fricatives", ICPhs XVII, pp. 663–666, 2011.

[13] D. Donnelly, C. Llamas and D. Watt, "Effects of different types of face covering on speech acoustics and intelligibility: some preliminary observations", IAFPA, 2007.

[14] 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", Interspeech, ISCA's 13th Annual Conference, Portland, OR, USA, pp. 2250–2053, 2012.

[15] J. Saigusa, "The Effects of Forensically Relevant Face Coverings on the Acoustic Properties of Fricatives", Lifespans and Styles, vol. 3, no. 2, pp. 40–52, 2017.

[16] A. J. Palmiero, D. Symons, Judge 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, n°12, pp. 960–968, 2016.

[17] X. Tang and X. Yan, "Acoustic energy absorption properties of fibrous materials: a review", Applied Science and Manufacturing, vol. 101, pp. 360–380, 2017.

[18] A. Chan, E. Gouva, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines and M. Seltzer, "The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Relate Resources", www.cs.cmu.edu/~archan/share/sphinxDoc.pdf, 2007.

[19] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvea, S. Radomski, M. Mühlhäuser and C. Biemann, "Open Source German Distant Speech Recognition: Corpus and Acoustic Model", Proceedings of Text, Speech and Dialogue (TSD), pp. 480–488, 2015.

[20] https://github.com/MaskSorbonneSpeechCorpus/sentences

[21] F. Eyben, F. Weninger, F. Groß and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in Proceedings of ACM MM, Barcelona, Spain, pp. 835–838, 2013.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel and J. Vanderplas, "Scikit-learn: Machine learning in Python", Journal of machine learning research, pp. 2825–2830, 2011.

[23] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates", in Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, ACM, pp. 694–699, 2002.

[24] F. Eyben, "Standard Baseline Feature Sets. In Real-time Speech and Music Classification by Large Audio Feature Space Extraction", Springer International Publishing, pp. 123–137, 2016.

[25] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, Prentice-Hall, 1989.

[26] N. Miller, "Measuring up to speech intelligibility", International Journal of Language and Communication Desorder, vol. 48, no. 6, pp. 601–612, 2013.

[27] T. R. Letowski and A. A. Scharine, "Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission", No. ARL-TR-8227, US Army Research Laboratory Aberdeen Proving Ground United States, 52 pages, 2017.

[28] K. D. Kryter and E. C. Whitman, "Some Comparisons between Rhyme and PB Word Intelligibility Tests", The Journal of the Acoustical Society of America, vol. 37, no. 6, p. 1146, 1965.

[29] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools", IEEE signal processing magazine, vol. 32, no. 2, pp. 114–124, 2015.

[30] K. M. Coyne and D. J. Barker, "Speech Intelligibility While Wearing Full-Facepiece Air-Purifying Respirators", Journal of occupational and environmental hygiene, vol. 11, no. 11, pp. 751–756, 2014.

[31] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem", Journal of the ACM (JACM), vol. 21, no. 1, pp. 168–173, 1974.

[32] J. Ma, Y. Hu and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", The Journal of the Acoustical Society of America, vol. 125, no. 5, pp. 3387–3405, 2009.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4214–4217, 2010.