INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Modeling the effect of military oxygen masks on speech characteristics

Benjamin Elie[1], *Jodie Gauvain*[2], *Jean-Luc Gauvain*[1], *Lori Lamel*[1]

[1]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
[2]Vocapia Research, Orsay, France

elie@limsi.fr, jodie@vocapia.com, gauvain@limsi.fr, lamel@limsi.fr

Abstract

Wearing an oxygen mask changes the speech production of speakers. It indeed modifies the vocal apparatus and perturbs the articulatory movements of the speaker. This paper studies the impact of the oxygen mask of military aircraft pilots on formant trajectories, both dynamically (variations of the formants at a utterance level) and globally (mean value at the utterance level) for 12 speakers.

A comparative analysis of speech collected with and without an oxygen mask shows that the mask has a significant impact on the formant trajectories, both on the mean values and on the formant variations at the utterance level. This impact is strongly dependent on the speaker and also on the mask model. These observations suggest that the articulatory movements of the speaker are modified by the presence of the mask.

These observations are validated via a preliminary ASR experiment that uses a data augmentation technique based on articulatory perturbations that are driven by our experimental observations.

Index Terms: speech variation, articulatory perturbation, oxygen mask, data augmentation

1. Introduction

The speech production of military aircraft pilots is affected by various environmental factors, such as loud noise and its subsequential effect, the Lombard speech [1, 2], strong acceleration (g-force) [3, 4, 5], the wearing of an oxygen mask [6, 7], and intense psychological workloads [8, 9].

Among these factors, the effect of wearing an oxygen mask has received attention since the 80's. Wearing an oxygen mask affects the speech production, since it modifies the vocal apparatus and perturbs the articulatory movements of the speaker. The vocal tract configuration is modified as it is closed by a chamber between the mask wall and the speaker's mouth. Speakers have reported [3] that the presence of the mask hinders their lips and jaw articulatory movements, which could result in a smaller mouth opening, or a limited lip protrusion [5]. It also increases the breath noise and adds various acoustical noises (e.g. respiratory valves, microphone artifacts). Bond *et al.* [6] analyzed acoustic-phonetic characteristics of speech wearing a mask and found that the overall effect of the oxygen mask is to compress the $F1 - F2$ vowel space. For other phonetic characteristics, such as fundamental frequency, and phoneme duration, wearing a mask had only a marginal influence. Other studies have investigated the impact of cognitive workloads on aircraft pilots and, consequently, analyzed speech of speakers wearing an oxygen mask. Keränen *et al.* [8] reported that the average fundamental frequency is raised by about 20% relative in a noisy condition. Huttunen *et al.* [9] found additional formants at around 300 Hz and between 2500 and 3000 Hz. Using numerical simulations of the transfer functions of the vocal tract to which an additional cavity is connected downstream, Vojnović *et al.* [10] associated the additional formant at 3000 Hz to the acoustic resonance of the mask cavity. They also observed a slight rise of the vocal tract resonances by at most 5%.

All of these speech modifications drastically degrade the performance of automatic speech recognition (ASR) systems when the speaker wears an oxygen mask [11]. Using recent speech recognition systems trained with normal speech [12, 13, 14], the Word Error Rate (WER) obtained for speech with the oxygen mask doubles in comparison to that of normal speech from the same speaker. In order to build accurate ASR models for military aircraft pilots, the speech variations needs to be clearly identified and quantified. One difficulty of the automatic recognition of military pilot's speech is the lack of available data (e.g. because of confidentiality issues). A possible solution could consist in artificially generating data by transforming normal speech so that it reflects the acoustic parameters of speech of military aircraft pilots. This approach can complement data augmentation techniques [15], such as the *Vocal Tract Length Perturbation* (VTLP) [16, 17], which is commonly applied to generate new data by virtually modifying the length of the vocal tract. Although this technique works well for generic tasks, we believe that it needs to be adapted to generate new data that fits the speech corresponding to that of the recognition task considered in this work.

For that purpose, this paper focuses on the analysis of the dynamic speech variations with an oxygen mask, with the aim of building transformation operators that can be applied to normal speech so that the transformed speech signal resembles that of speech with an oxygen mask. Acoustic variations of speech are analyzed at the utterance level with a focus on the articulatory perturbations due to the presence of the mask by comparing the same sentences pronounced by the same speakers both with and without the mask. The experimental study analyzes the global and local variations of formant trajectories in speech from 12 speakers. In comparison to previous studies (e.g. [3]), variations of formant values are analyzed both dynamically (variations of the formants at a utterance level) and globally (mean value at the utterance level). Additionally, individual variations of speech across speakers are analyzed. In order to validate our model, preliminary ASR experiments are carried out using an articulatory-based perturbation technique for data augmentation, where the articulatory perturbations are driven by our experimental observations.

The organization of this paper is as follows. Section 2 describes the experimental observation of the speech variations with an oxygen mask, including the subjects, the analysis methods, and the presentation of the results. Section 4 presents the ASR experiments using an articulatory-based data augmentation technique to improve the performance of ASR tasks on

http://dx.doi.org/10.21437/Interspeech.2021-1650

speech with an oxygen mask.

2. Speech corpora

This section describes the corpora used for the two parts of the paper: the experimental analysis of the effect of the oxygen mask on speech characteristics and the preliminary ASR experiment. The corpus used for the analysis of speech variations is designed to be easy to utter: a dozen of syllables per sentence and all sentences are in the native language of the speakers (i.e. French) with a close to standard phraseology. The corpus for the ASR experiments was designed to target pilot's speech, namely with the specific Air Traffic Control and military aviation grammar phraseology.

2.1. Corpus used for the analysis of speech variations

Data were collected in two recording sessions. The first one includes 8 subjects, 7 male speakers (BE, CB, JD, MCB, MV, PYL, and WB), and 1 female speaker (JG). The second one includes 6 subjects, 2 male speakers (BE and MS), and 4 female speakers (JG, FSA, FSB, and FSC). Speakers BE and JG participated in both sessions. All 12 subjects are French native adult speakers with no known speech disorder or hearing impairment.

The text corpus used for both recording sessions consists of 50 sentences taken from the Combescure corpus [18]. This corpus is a well know reference for linguistic studies and was specifically designed to reproduce all French phonemes with individual recurrence rates similar to those encountered in spoken French. The choice of this corpus was also motivated by the fact that the sentences are short enough to limit pronunciation errors and respiratory pauses when using the oxygen mask, while being long enough to observe the acoustic effects of the articulatory gestures. The shortest sentence has 8 syllables and the longest 18.

Recordings were made in a quiet environment. The no-mask recordings were acquired with a dynamic headset microphone and using the capsule of a dynamic microphone mounted in the oxygen mask for the mask recordings. Data were sampled at the rate of 8 kHz for the post-processing. Two kinds of oxygen masks were used for the experiment. The first one (denoted M1) is a Ulmer 82AB and the second one (denoted M2) is a Ulmer UA21S.

2.2. Corpus used for the ASR experiment

Since the aim of the prelimary ASR experiment is to validate the proposed approach, a small, task-specific system was constructed. The training data come from two corpora in English totaling 15.7 hours of speech: (1) Air Traffic Control (ATC) data (2) recordings of read speech using the official military aircraft phraseology. The ATC corpus contains 7000 utterances with a total duration of 8.1 hours and the phraseology corpus contains 7628 utterances with a total duration of 7.6 hours.

The test corpus is a set of 385 sentences with a mix of official military aircraft phraseology and sentences inspired from real flight missions. All sentences were uttered by 4 different speakers with an oxygen mask in quiet conditions, for at total duration of 18 minutes.

3. Analysis of speech variations

Formant trajectories were extracted from audio recordings using the Parselmouth library [19], which is a Python interface to access Praat functionalities [20]. From the trajectories of the first three formants at the utterance level, two parameters were computed for each individual trajectory: the median formant $\bar{F}i$ and the formant span Si.

The formant span computation is an adaptation of the method initially proposed by Jany-Luig [21] for pitch contours, but applied to formant trajectories. It is based on previous work by Patterson and Ladd [22], in which the authors proposed to use clear pitch marks (maximums and minimums of the pitch contour) to compute the pitch span. Then, the authors defined the pitch span as the difference between the average distance between the maximums and the local mean contour and the average distance between the minimums and the local mean contour. By applying this method to the formant trajectories, it is possible to get a feature that measures the range of articulatory movements produced by the subject: for a given utterance, the lower the formant span, the smaller the articulatory gesture. The median formant is defined as the median value of an individual formant across the whole utterance.

For each parameter, we computed a normalized value, named the R-index, defined as the value for a given sentence uttered with the oxygen mask normalized by the value of the same sentence uttered by the same speaker without the mask. Hence a value larger than 1 means that this parameter is raised in the mask condition, and a value smaller than 1 means that it has been reduced. We also performed a one-way ANOVA on the two groups of data (without and with a mask) to assess the significance of the effect of the mask for each parameter. Another one-way ANOVA is performed for the relative data considering each speaker-session as a separate group.

Results are presented for each subject and experimental session. The label indicates the following information: subject's initials, session number (1 or 2), gender (M or F), mask number (1 or 2). For instance, BE1M1 means that the data corresponds to subject BE, recorded in the first experimental session, the subject is male and wearing the mask M1.

3.1. Median formant

Table 1: *Median values of R_{Fi} for the first three formants for each recording session. Numbers in brackets are the standard variations. Scores in blue font are less than 1. Scores followed by $*$ are statistically significant with $p < 10^{-5}$.*

	R_{F1}	R_{F2}	R_{F3}
BE1M1	1.05 (0.06)	0.98 (0.06)	0.95 (0.05)*
BE2M2	0.88 (0.04)*	0.92 (0.07)*	0.87 (0.03)*
CB1M1	1.07 (0.06)*	0.99 (0.06)	0.98 (0.03)
FSA2F1	1.14 (0.06)*	1.04 (0.08)	0.94 (0.04)*
FSB2F1	1.27 (0.13)*	1.03 (0.08)	0.94 (0.05)*
FSC2F1	1.26 (0.11)*	1.09 (0.10)*	0.92 (0.05)*
JD1M1	1.04 (0.08)	0.97 (0.11)	1.03 (0.06)
JG1F1	0.96 (0.07)	0.83 (0.08)*	0.94 (0.04)*
JG2F2	0.85 (0.08)*	0.79 (0.07)*	0.91 (0.03)*
MCB1M1	0.98 (0.11)	0.96 (0.09)	0.98 (0.03)
MS2M2	0.94 (0.07)	0.89 (0.07)*	0.90 (0.03)*
MV1M1	1.00 (0.08)	0.91 (0.11)	0.98 (0.05)
PYL1M1	1.03 (0.08)	0.90 (0.08)*	0.96 (0.05)
WB1M1	0.93 (0.07)*	0.78 (0.16)*	1.04 (0.04)*
all speakers	1.00 (0.15)	0.94 (0.12)*	0.94 (0.06)*

Table 1 displays the values of R_{Fi}, i.e. the relative modifications of the median formant frequency in the mask condition

compared to the no-mask condition, for the first three formants $F1$, $F2$, and $F3$. It shows a large cross-speaker variability, with no clear general tendency across speakers. For instance, some speakers raise the first formant $F1$ (BE with mask M1, CB, FSA, FSB and FSC), while it is lowered for others (BE with mask M2, JG with mask M2, and WB1). The modification may be large : values range from -15% upto $+26\%$. Grouping the data from all speakers leads to a median variation of -0.25% ($p = 0.66$). The results for BE and JG show that there is a combined effect of the type of mask and its placement on the speaker: BE raises $F1$ with M1 but lowers it with $M2$, and the lowering of $F2$ for JG is much larger with M2 (-15%) than with M1 (-4%).

The second formant $F2$ shows less variation than $F1$ across speakers. Globally, speakers tend to lower $F2$, with a median variation of -6% across all speakers. The decrease is almost -20% for JG with mask M2. However, the presence of the oxygen mask has no effect on the $F2$ for some speakers (BE for mask M1, CB, FSA, FSB, and JD), and for one speaker, FSC, $F2$ raises with a median value of $+9\%$.

The results for the third formant $F3$ are similar to those observed for $F2$: there is a global tendency for speakers to lower $F3$. The median variation across speakers is -6.2%. The median value of R_{F3} of individual speakers is generally between -5% and -15%.

Overall it can be seen that the formant values are highly impacted by the presence of the oxygen mask. While there is no clear tendency for $F1$, which can be either raised or lowered, depending on the speaker, $F2$ and $F3$ are usually lowered by usually less than 20%.

3.2. Formant span

The relative modifications of the formant span, analyzed through R_{Si} are displayed in Table 2. It shows that the median value of R_{S1} is less than 1 for most speakers. The median ratio across all speakers is 0.73, which corresponds to a lowering of about 27%. The minimal median R-index for the $S1$ is -55% for WB. Only speaker MS has an R_{S1} mostly larger than 1. Since $F1$ is commonly associated with the jaw and lip movements, this suggests that the speakers tend to decrease their jaw movements when wearing an oxygen mask. This is in agreement with remarks made by pilots that have been reported in previous studies [3, 5].

The median value of R_{S1} across all speakers is -4%, with $p = 0.54$. However, a large cross-speaker variability is observed. Indeed $S2$ is raised significantly for CB, JG with mask M1, MCB, MV, and WB. On the contrary, it is significantly reduced for speakers BE and JG, both with mask M2, FSA, FSB, and FSC. There is no significant impact for other speakers ($p > 0.1$). Since, the second formant frequency is commonly associated with the tongue position, one hypothesis could be an articulatory compensation, but further experiments should be carried out to validate it.

Finally, R_{S3} also varies significantly according to the speaker. Across all speakers, the median value is $+0.99\%$ ($p = 0.001$), but it is significantly increased for speakers BE, MCB, MS, MV, and PYL, with median values going from $+18.7\%$ to $+49.0\%$ ($p < 0.007$), and decreased for speakers FSA, FSB, FSC, and WB, with median values going from -68.4% to -27.0%.

Table 2: *Median values of R_{Si} for the first three formants for each recording session. Numbers in brackets are the standard variations. Scores in blue font are less than 1. Scores followed by $*$ are statistically significant with $p < 10^{-5}$.*

	R_{S1}	R_{S2}	R_{S3}
BE1M1	0.72 (0.21)*	1.12 (0.43)	1.32 (0.48)*
BE2M2	0.66 (0.19)*	0.77 (0.27)*	1.19 (0.46)
CB1M1	0.82 (0.33)	1.25 (0.30)	1.02 (0.27)
FSA2F1	0.72 (0.19)*	0.73 (0.17)*	0.66 (0.14)*
FSB2F1	0.94 (0.36)	0.76 (0.16)*	0.44 (0.13)*
FSC2F1	0.78 (0.26)	0.54 (0.14)*	0.32 (0.11)*
JD1M1	0.73 (0.22)	1.04 (0.44)	0.84 (0.57)
JG1F1	0.67 (0.28)*	1.48 (0.51)*	0.97 (0.32)
JG2F2	0.89 (0.35)	0.71 (0.32)*	1.01 (0.29)
MCB1M1	0.49 (0.24)*	1.36 (0.54)*	1.25 (0.45)
MS2M2	1.31 (0.50)*	0.99 (0.32)	1.49 (0.47)*
MV1M1	0.73 (0.31)*	1.31 (0.41)*	1.27 (0.34)
PYL1M1	0.70 (0.29)*	1.09 (0.41)	1.37 (0.38)*
WB1M1	0.45 (0.22)*	1.73 (0.50)*	0.73 (0.21)*
all speakers	0.73 (0.36)*	0.98 (0.49)	1.01 (0.49)

4. Articulatory-based perturbations for data augmentation

This section presents some preliminary speech recognition experiments using an articulatory-based data augmentation technique. These experiments make use of task-specific data for model training and an independent test corpus. The data augmentation technique is based on articulatory perturbations that model the effect of the oxygen mask on the formant trajectories that were presented in Section 3.

4.1. Articulatory perturbations

From the data presented in Section 3, it is possible to extract (non-parametric) probability density functions for each formant parameter and each subject. Then, for any utterance in the initial dataset, a new value of the formant parameters is randomly chosen from the associated probability density function, and data augmentation is performed by applying the appropriate transformations to the formant trajectories. Changing the formant median value F_i by a random value α_i is associated with a translation, namely $F_i' = \alpha_i F_i$. Scaling the formant span by a random value β_i, requires applying a detrending operator $\mathcal{L}(\mathbf{F}_i)$, which consists of subtracting the falling or rising line of declination of the formant trajectory $\mathbf{F}_i$ and subsequently centering it around zero by subtracting the mean. The inverse operation is then denoted $\mathcal{L}^{-1}(\mathbf{F}_i)$. The new formant trajectory is then $\mathbf{F}_i' = \mathcal{L}^{-1}\left[\beta_i \mathcal{L}(\mathbf{F}_i)\right]$.

Finally, each windowed signal frame $x_w(t)$ of the audio speech signal is transformed using a piecewise linear frequency warping function on the spectral envelope. The transformation acts such that the peaks of the intial spectral envelope corresponding to the initial formant frequencies F_i match the target formant frequencies F_i'. The new signal frame $x_w'(t)$ is then

$$x_w'(t) = \mathcal{F}^{-1}\left[X_w(f)\frac{|C_x(f')|}{|C_x(f)|}\right], \qquad (1)$$

where $X_w(f) = \mathcal{F}\left[x_w(t)\right]$ is the Fourier transform of the windowed signal frame $x_w(t)$, the operator $\mathcal{F}^{-1}$ is the inverse Fourier transform, $C_x(f)$ is the cepstrum-based spectral enve-

lope of $x_w(t)$, and $C_x(f')$ is the frequency-warped version of $C_x(f)$. Figure 1 illustrates the transformation on the formant trajectories and the piecewise linear frequency warping function applied to the spectral envelope.

Audio transformations were performed using the articulatory perturbations application of the SpeechHiker Python library[1].

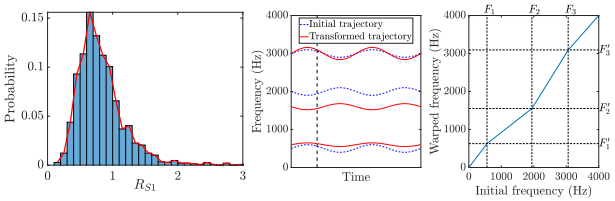

Figure 1: *Example transformations applied to the formant trajectories and the piecewise linear frequency warping function applied to the spectral envelope. Left: transformation factors α and β are randomly generated following a probability density function computed from our statistical observations (here R_{S_1} for all speakers). Center : the initial formant trajectories for $F1$, $F2$, and $F3$, and the trajectories resulting from the transformation factors of $\alpha = [1.2, 0.8, 1]$ and $\beta = [0.4, 1, 1.6]$. Right: piecewise linear frequency frequency warping function at the time frame represented by the dashed vertical line in the left plot.*

4.2. Preliminary speech recognition experiments

Table 1 gives the word error rate (WER) obtained on the test data using a speech recognizer [14] relying on a HMM-TDNN [23] acoustic model and a standard 3-gram language model. The baseline training result is reported in the first row of the table, followed by the result using a classical vocal tract length perturbation (VTLP) data augmentation method, and using the proposed articulatory perturbation method based on the statistical distribution of the articulatory variations taken across all speakers. Since we are making use of a non-deterministic NN training process, all models have been trained five times and the WERs averaged.

Table 3: *WER obtained on the test data with ASR models trained with normal speech (baseline), adding VTLP-based augmented data (+VTLP), and adding the data augmented using our articulatory perturbation technique following the statistical distribution obtained from all speakers (+AP(all speakers)).*

Training	# hours	mean WER (%)
Baseline	15.7	25.7
+VTLP	31.4	21.6
+AP(all speakers)	31.4	20.0

The models trained with the augmented corpus are seen to outperform the baseline models. The best performance is obtained using our novel articulatory-based augmentation technique. The WER is reduced by over 20% (max is 22%) from the baseline value (20.0% vs. 25.7%). The articulatory-based augmentation technique also outperforms the VTLP-based augmentation technique, which reduces the WER by 16% from the baseline.

[1] https://gitlab.com/benjamin.elie/spychhiker

5. Discussion and conclusions

This paper has presented a study about the variations of speech production when wearing a military aircraft oxygen mask. It consists of a comparative experimental study that analyzes the impact of the oxygen mask on articulatory features extracted from real speech recordings. Our experiments reported large cross-speaker variability for these features that have not been reported in previous studies [3, 9]. Given the hypothesis that the mask perturbs the articulatory movements of the speakers, the cross-speaker variability can be explained by several factors such as the speaker physiology and the fit and degree of mask tightening. Despite the large variability, the experiments highlight global tendencies.

The formant trajectories are significantly impacted by the presence of the oxygen mask. Indeed, compared to normal speech, the speakers' mean formant frequencies are modified. For $F1$, speakers either increase or decrease the mean formant frequency, from -15% upto $+26\%$. $F2$ and $F3$ are generally reduced by 20% and 15%, respectively. The global shape of the formant trajectories is also modified by the presence of the oxygen mask. The most significant variation lies in $F1$, for which our experiments show an important reduction of the variations of the trajectory around the mean value. This reduction of the formant span is between 20% and 40% among speakers, compared to the formant span observed in normal speech. These results support the hypothesis of articulatory perturbations due to the presence of the mask, especially for the jaw and the lips, which are commonly associated with $F1$ variations. This hypothesis is also supported by the fact that previous studies showed that adding a cavity downstream the vocal tract has only a minor influence on formants [24, 10, 25].

This paper also presented a preliminary ASR study which uses an articulatory-based data augmentation technique to train models with data that reflect the aforementioned observed acoustic variations. This data augmentation technique extends the VTLP technique by a piecewise linear frequency warping function applied to the spectral envelope that enables each formant to be modified independently at each time frame. Thus, it is possible to change the formant trajectories of the original speech so that they correspond to typical formant trajectories observed in speech recorded wearing an oxygen mask. These preliminary ASR results validate our proposed method: in comparison with models trained with the original speech data, the WER of the test data is reduced when using models trained with the articulatory-based augmented data. This shows that specific ASR tasks can benefit from adapted data augmentation techniques.

In this paper, we only focused on articulatory perturbations because they are likely the most significant factor explaining the acoustic variations of speech with an oxygen mask. However, other factors that may be taken into account in future work include microphone effects and prosodic variations due to cognitive workloads [8, 9].

6. Acknowledgements

This study is partly supported by the Man Machine Teaming project MEVAC and the RAPID project both funded by the Defence Innovation Agency (Agence de l'innovation de défense) in liaison with the French Armement General Directorate (Direction générale de l'armement). The authors would also like to thank Synapse Défense and the CEAM for having participated to the recording sessions.

7. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Mal. de L'Oreille et du Larynx*, pp. 101–119, 1911.

[2] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[3] Z. S. Bond, T. J. Moore, and T. R. Anderson, "The Effects of High Sustained Acceleration on the Acoustic Phonetic Structure of Speech. A Preliminary Investigation," tech. rep., 1986.

[4] C. Gulli, D. Pastor, A. Leger, P. Sandor, J. Ciere, and P. Grateau, "G-load effects and efficient acoustic parameters for robust speaker recognition," *Advanced Aircraft Interfaces: the Machine side of the Man-Machine Interface*, 1992.

[5] A. J. South, "Some characteristics of speech produced under high G-force and pressure breathing," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 4, pp. 2095–2098, IEEE, 1999.

[6] Z. Bond, T. J. Moore, and B. Gable, "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 907–912, 1989.

[7] M. Vojnović and M. Mijić, "The influence of the oxygen mask on long-time spectra of continuous speech," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2456–2458, 1997.

[8] H. Keränen, E. Väyrynen, R. Pääkkönen, T. Leino, P. Kuronen, J. Toivanen, and T. Seppänen, "Prosodic features of speech produced by military pilots during demanding tasks," in *Proceedings of the Fonetiikan Paivat Conference*, pp. 88–91, 2004.

[9] K. H. Huttunen, H. I. Keränen, R. J. Pääkkönen, R. Päivikki Eskelinen-Rönkä, and T. K. Leino, "Effect of cognitive load on articulation rate and formant frequencies during simulator flights," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1580–1593, 2011.

[10] M. Vojnović, M. Mijić, and D. Šumarac-Pavlović, "Transfer characteristics of vocal tract closed by mask cavity," *Archives of Acoustics*, vol. 43, 2018.

[11] H. J. Steeneken and J. v. Velden, "The effect of an oxygen mask on automatic speech recognition," in *Applications of Speech Technology*, 1993.

[12] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1, pp. I–I, IEEE, 2003.

[13] A. Laurent, T. Fraga-Silva, L. Lamel, and J.-L. Gauvain, "Investigating techniques for low resource conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5975–5979, IEEE, 2016.

[14] J. Gauvain, L. Lamel, V. B. Le, J. Despres, J.-L. Gauvain, A. Messaoudi, B. Vieru, and W. B. Kheder, "Challenges in audio processing of terrorist-related data," in *International Conference on Multimedia Modeling*, pp. 80–92, Springer, 2019.

[15] J. M. Ramirez, A. Montalvo, and J. R. Calvo, "A survey of the effects of data augmentation for automatic speech recognition systems," in *Iberoamerican Congress on Pattern Recognition*, pp. 669–678, Springer, 2019.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[17] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," *INTERSPEECH-2019, Graz, Austria*, pp. 739–743, 2019.

[18] P. Combescure *et al.*, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 34–38, 1981.

[19] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," 2018.

[21] J. Jany-Luig, *Prosodic and Paralinguistic Speech Parameters for the Identification of Emotions and Stress*. PhD thesis, Faculty of Psychology and Neuroscience, Maastricht University, 2017.

[22] D. Patterson and D. R. Ladd, "Pitch range modelling: linguistic dimensions of variation," in *Proceedings of ICPhS*, vol. 99, pp. 1169–1172, 1999.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 3214–3218, 2015.

[24] A. J. South, "A model of vowel production under positive pressure breathing," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[25] M. Vojnović, M. Mijić, D. Šuramac-Pavlović, and N. Vojnović, "Influence of overpressure breathing on vowel formant frequencies," *Archives of Acoustics*, vol. 46, no. 1, pp. 177–181, 2021.