VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2021.2.14
Vibrations in Physical Systems, 32(2):2021214, 2021

Effect of Sleepiness in the Voice on Speaker Recognition Performance

Piotr STARONIEWICZ

Corresponding author: Piotr STARONIEWICZ, email: piotr.staroniewicz@pwr.edu.pl

Wroclaw University of Science and Technology, Department of Acoustics, Multimedia and Signal Processing, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

Abstract The issue of the influence of speaker state on voice recognition has been analysed mainly in relation to forensics and biometric security systems. Sleepiness in the voice is a rather under-researched problem, and the few works in this area focus almost exclusively on the recognition of sleepiness rather than on its influence on the change of the speaker's voice characteristics. This paper discusses the issue of the influence of the speaker's state on voice recognition, describes the acquisition method of the acoustic database of voice drowsiness recordings used in the tests. It also discusses the subjective sleepiness scales used in the study and presents the results of the influence of sleepiness on the effectiveness of automatic speaker recognition based on a classical system using the Mel-Frequency Cepstral Coefficients parameterisation and the Gaussian Mixture Models classification.

Keywords: speaker recognition, sleepiness

1. Introduction

Increasingly effective biometric systems that recognise the voice of the speaker are now being used very widely for access control security, transaction authentication and also in forensics. Regardless of whether voice recognition is performed by a human (e.g. a phonoscopy expert) or, as it is increasingly the case, by an automated system, the measurable parameters of a speaker's voice can change significantly. These changes are due to the fact that the human voice, unlike such biometric traits as iris, DNA or even fingerprints, which are fairly constant over time, the human voice is subject to numerous noticeable changes due to ageing, emotions, health and many other factors. Since these factors, whatever their origin, can be regarded as a deviation, transformation or distortion of a 'normal voice', they are treated as voice disguises.

A factor that affects the functioning of our body (from cognitive abilities, to slowing down psychomotor reactions, to spatial orientation or decision-making abilities) is sleepiness or fatigue. Therefore, its detection and assessment is of interest to researchers and has applications in road traffic (e.g. assessment of drivers' state), safety-relevant facilities (e.g. chemical plants, nuclear power plants, air traffic). Relatively few works on the issue of sleepiness analysis in the voice focus almost exclusively on the problem of its detection. However, an important aspect remains unexplored, namely the influence of changes in voice parameters due to fatigue and sleepiness on the efficiency of speaker recognition.

2. Influence of speaker's state on voice recognition

The voice disguises can be classified according to two independent divisions: as deliberate – nondeliberate (intentional – unintentional), and as technical – natural. The voice disguises caused by a change in speaker's state can be classified as non-deliberate natural ones. Table 1 summarises the types and examples of such disguises [1–4].

One of the most important types is ageing - continuously occurring anatomical and physical changes manifested in the produced speech by: voice tremor, slower articulation, laryngeal tension and air loss. The effect of intoxication (i.e. alcohol or drugs) of subjects on speech depends largely on individual characteristics and the amount of the substance in the speaker's blood, and manifests itself most commonly in speech by: slowing of speaking rate and changes of the fundamental frequency distribution. The emotional state of the speaker is another important factor influencing the speech parameters, especially changes in the value and waveform of the fundamental frequency of the laryngeal tone. Voice parameters

are often used to identify selected basic emotional states (e.g. the so-called "Big Six", i.e.: anger, disgust, fear, happiness, sadness and surprise) [5].

Diseases such as hoarseness, laryngitis and pathological changes in the speech organs can also significantly affect the spectral characteristics of the speech signal. This is used in acoustic medical diagnostics and is also reflected in the deterioration of the performance of voice recognition systems. The last non-deliberate natural disguise category proposed in the taxonomy presented in Table 1 is the change in speaker's conditions. This category refers to factors not mentioned earlier which also affect the voice of the speaker. This includes the influence of external factors on the psychophysical functioning of the speaker's body, such as loud noise (Lombard effect), pain, high or low temperature, vibration etc. This category also includes factors such as sleepiness or fatigue of the speaker.

Tab. 1. Types of non-deliberate natural voice disguises caused by change in speaker's state.

Types of change in speaker's state	Examples
Aging	anatomical and physical changes occurring naturally during life
Intoxication	articulation under the influence of drugs or alcohol
Emotional state	speech affected by emotions
Illness	diseases affecting the speech organ (e.g. hoarseness or laryngitis)
Change in condition	sleepiness, fatigue or impact of external conditions

Sleepiness in the voice is still at present a rather poorly studied problem. In fact, all the relatively rare papers in this field focus almost exclusively on the recognition of drowsiness rather than on its effect on the change in the personal characteristics of the speaker's voice [6,7]. Table 2 groups the most significant changes occurring in the human body as a result of sleepiness into five categories and lists how they may affect voice parameters.

Tab. 2. Drowsiness-induced changes in the human body that may affect the characteristics of speech parameters.

Types of change	Alterations in the parameters of the voice in relation to natural speech
Reduced cognitive speech planning	Slacked articulation and slowed speech
Flatter and slower respiration	Lower fundamental frequency, intensity and rate of speech
Reduction in vocal fold tension	Spectrum energy shift, decreased formants' values and positions
Softened vocal tract walls and pharynx	Wider formant bandwidth
Reduced mobility of the orofacial region and mouth	Slacked articulation, increased nasality

Due to its highly subjective nature, a difficult yet important issue in all studies conducted on sleepiness is how to evaluate its degree. Depending on the purpose, different subjective scales are used in sleepiness studies to allow for self-assessment [8]. The Epworth Sleepiness Scale (ESS) is used in the assessment of sleep disorders and is based on the respondent's determination (on a scale of 0-3) of the likelihood of falling asleep in eight everyday situations. The Karolinska Sleepiness Scale (KSS) measures subjective levels of sleepiness at a specific time during the day. It is a nine-point scale (in addition to the classic scale, there are also its numerous modifications, which are often ten-point scales or even more) in which respondents indicate the level that best reflects their psycho-physical state in the last 10 minutes. It is a measure of

situational sleepiness that is sensitive to fluctuations [8]. Stanford Sleepiness Scale (SSS) is a subjective, seven-item measure of sleepiness. It is used for both research and clinical purposes. While other tests take into account the general feeling of sleepiness throughout the day, the SSS scale allows the assessment of sleepiness at a particular moment in time, hence it is well suited for repeated use during the course of a study. The SSS is freely available online.

3. Database

The acoustic base used in the tests was recorded at hourly intervals throughout the night. Once the participants had completed the form, the recordings began and lasted for 8 hours from 8 p.m. to 4 a.m. Starting at 8 p.m., every hour each participant was asked to rate subjectively their perceived level of sleepiness, and after the rating, a voice sample was recorded, consisting of saying the vowel 'a' 3 times and saying the sentence in Polish 'Ala ma kota' (Sampa: 'ala ma kota' / in English: 'Ala has a cat') 3 times. The sentence had been chosen to be short, easy to remember and pronounce without requiring the speaker to be very focused during articulation (it is usually the first sentence in a primer when learning to read). This sentence also contains a high content of voiced phonemes to facilitate the recognition of the speaker's voice. Each time, every hour, immediately after the recording, each speaker completed a self-assessment according to the seven-point Stanford Sleepiness Scale (SSS) learned prior to the session. The subjective scores obtained on the SSS scale for each speaker are summarised in Tab. 3.

Tab. 3. Hourly subjective evaluation of speakers according to the SSS scale.

Speaker			Recording's Hour								
Number	Male/Female	Age	20.00	21.00	22.00	23.00	00.00	01.00	02.00	03.00	04.00
1	F	21	2	2	2	2	5	5	6	7	7
2	F	24	1	1	3	4	5	6	6	7	7
3	F	45	2	2	4	5	5	5	6	6	7
4	M	23	2	2	2	3	5	5	5	7	7
5	M	29	2	2	4	4	5	5	6	6	7
6	M	44	1	2	3	3	3	4	6	6	7

Significant changes in speech signal parameters that may occur in the speaker as a result of his or her fatigue and sleepiness are shown in Fig. 1. The shown sample waveforms of the fundamental frequency for a selected speaker indicate that there is, among other things, a decrease in the mean value and dynamic of the fundamental frequency. This can have a significant impact on the deterioration of speaker recognition performance.

4. Results and discussion

In speaker recognition systems based on Gaussian Mixture Models (GMM), the most common way to obtain a background model is to take speech samples from a group of speakers and train a single model. The model trained in this way is called the main, global or universal model. The main advantage of this method is that the background model can be trained once and then used each time to identify all the speakers, reducing both the verification time and the memory space required. This method has become the dominant method in speaker verification systems and is called GMM background model. The second method, referred to herein as the maximum likelihood method, requires the determination of an average or maximum value from a set of likelihoods of alternative speaker models. This method requires the system to create a separate background model for each speaker, which can be cumbersome in systems where the number of users is relatively large [10,11].

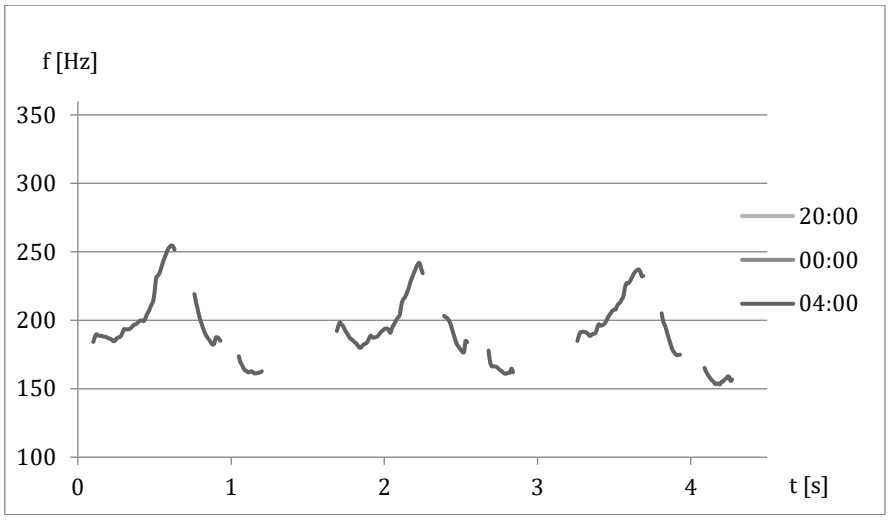

Fig. 1. Fundamental frequency waveforms for three recordings of speaker 2, with the same content, three repetitions of the utterance 'Ala ma kota', made at: 20.00, 00.00 and 04.00.

Two types of errors can occur in the speaker verification process: false acceptance rate (FAR) and false rejection rate (FRR), which are monotonic functions of the decision threshold. The pair of these errors determines the operating point of the verification system. The relationship between FAR and FRR on a Gaussian scale is referred to as detection error tradeoff (DET), while the operating point when both values are equal is called equal error rate (EER). This is used as a measure of the effectiveness of the verification system (the lower the EER the higher the effectiveness of the system) [12].

The window size mostly ranges from 20ms to 30ms. The best results were obtained for a window size of 1024 samples at a sampling rate of 44.100 kHz (23ms). The Hamming window and its modified version, the Hanning window, are the most commonly used in the speaker recognition process. In our case, the smallest EER error values were found for the Hamming window. 15 filters were used in the mel scale transition. The number of unimodal Gaussian distributions for the maximum likelihood method was chosen to be 24, while for the background model method it was 32 distributions.

The original recordings of each of the 20 o'clock speakers were taken as the learning sets. They last about 10 seconds (repetitions of vowels and sentences). The testing sets, on the other hand, last about 4 seconds and contain the sentence 'Ala ma kota' spoken three times. The test was divided into three parts. In the first part, recordings from 21.00 and 22.00 were used as test recordings, when the participants were not tired (depending on the speaker, a score of 1, 2 or 3 on the SSS scale). The second part contains recordings from 23.00 to 01.00, for which there is slight drowsiness (score 3 to 6 on the SSS scale). The last section contains samples recorded for hours: 2.00, 3.00 and 4.00, when participants felt strong fatigue (a score of 6 or even 7 on the SSS scale).

The test results are presented in the form of DET characteristics in Fig. 2 (results for GMM maximum likelihood method) and Fig. 3 (results for GMM background model) and as EER values collected in Tab.4. Analysing the results obtained, it can be seen that for the maximum likelihood method there is the lowest percentage of misrecognitions and rejections. In the case of the studied speaker base, verification with the background model method is less effective, and even inadvisable due to the small number of people in the base. The most significant finding obtained from the tests is that the EER error increases with the increasing sleepiness. Consequently, the speaker recognition system is most effective in recognising speakers making the recordings for hours 21 and 22, the EER error in this case being 3.33%. For the recordings from hour 23 to hour 4, the EER error was equal to 11.11%. For the tests carried out using the background model, a similar relationship was observed, and thus for samples from 21 to 22, the EER error was 8.33%, (i.e. 5% more than for the maximum likelihood method verification), in the interval for recordings from 23 to 1 o'clock, the EER error was equal to 14.44% and the largest EER error for the background model method was 16.67% (i.e. 6% more than in the maximum likelihood method).

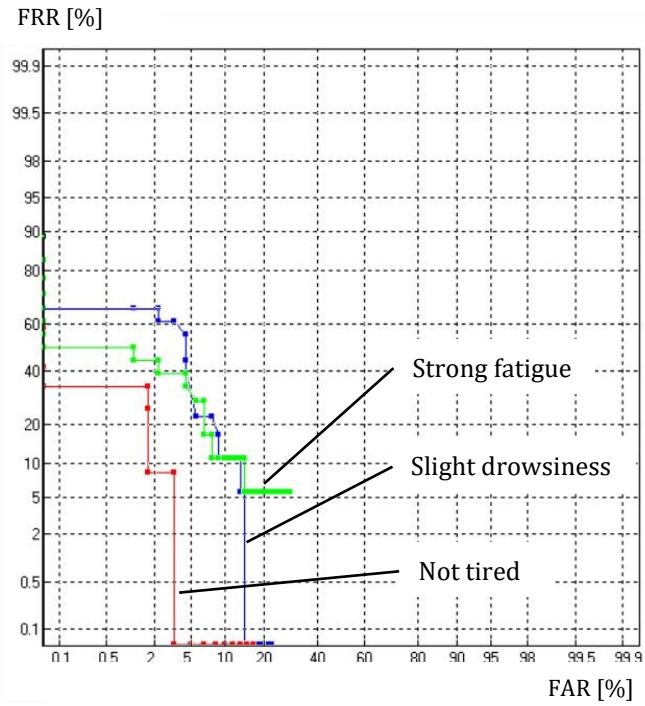

Fig. 2. DET curves for GMM maximum likelihood method.

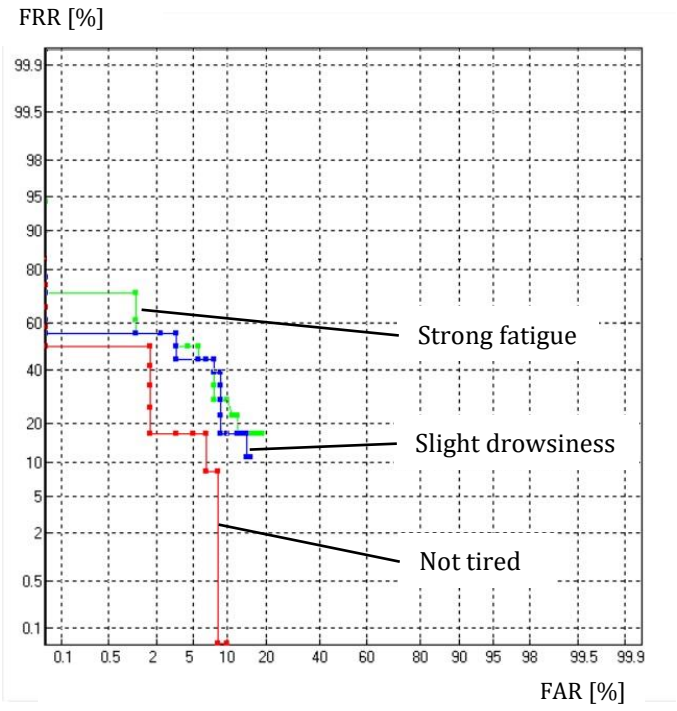

Fig. 3. DET curves for GMM background model method.

Tab. 4. EER rates.

	GMM maximum likelihood method	GMM background method
Not tired (21.00-22.00)	3,33 %	8,33 %
Slight drowsiness (23.00-01.00)	11,11 %	14,44 %
Strong fatigue (02.00-04.00)	11,11 %	16,67 %

5. Conclusions

The work carried out so far in the field of voice sleepiness research has focused exclusively on the aspect of detecting the speaker's sleepiness. The benefits of the potential application of such techniques for improving the control of people responsible for our safety (flight controllers, drivers, power plant operators, etc.) cannot be overestimated. However, a hitherto unexplored aspect has been neglected, namely the effect of masking the personal characteristics of the voice.

Due to the arduous conditions of the recordings carried out so far, tests have been conducted on a fairly limited number of voices, although it is planned to increase the size of the test base for further work. Despite this, the tests confirmed that voice fatigue and sleepiness are factors that should be taken into account when using speaker recognition systems.

In parallel to expanding the speaker database, future work in this aspect will focus on attempts to improve the robustness of speaker recognition systems by taking into account the influence of a person's state on their voice parameters.

References

1. C. Zhang, T. Tan. Voice disguise and automatic speaker recognition. Forensic Science 35 International, 175:118-122, 2008.
2. P. Staroniewicz. Effect of deliberate and non-deliberate natural voice disguise on speaker recognition performance. Acoustics, Acoustoelectronics and Electrical Engineering, 312-325, 2021.
3. M. Farrus. Voice Disguise in Automatic Speaker Recognition. ACM Computing Surveys, 51(4), 2018.
4. P. Staroniewicz. Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition. Proc. Of Joint Conference – Acoustics, IEEE, 2018.
5. P. Staroniewicz. Considering basic emotional state information in speaker verification. Proc. 4th International Conference on Biometrics and Forensics (IWBF) IEEE, 2016.
6. J. Krajewski, A. Batliner, M. Golz. Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern approach. Behavior Research Methods, 41(3):795-804, 2009.
7. J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller. Applying Multiple Classifiers and Non-Linear Dynamic Features for Detecting Sleepiness from Speech. Neurocomputing, 84:65-75, 2012.
8. A. Shahid et al. (eds.). STOP, THAT and One Hundred Other Sleep Scales. Springer Science+Business Media, 2012.
9. A. A. Miley, G. Kecklund, T. Akerstedt. Comparing two versions of the Karolinska Sleepiness Scale (KSS). Sleep Biol. Rhythms, 14:257-260, 2016.
10. D. A. Reynolds, R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech and Audio Proc., 3(1):72-83. 1995.
11. D. A. Reynolds, T. F. Quatieri, R. B. Dunn. Speaker Verification Using Adapted Gausian Mixture Models. Digital Signal Processing, 10:19-41, 2000.
12. A. Martin, A. Doddington, T. Kamm, M. Ordowski, M. Przybocki. The DET Curve in Assessment of Detection Task Performance, EuroSpeech 1997, Proceedings, 4:1895-1898, 1997.