IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Speaker Recognition under Stress Conditions

Esther Rituerto-González, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno

Signal Theory and Communications Department
University Carlos III Madrid

erituert@ing.uc3m.es, gallardo@tsc.uc3m.es, carmen@tsc.uc3m.es

Abstract

Speaker Recognition systems exhibit a decrease in performance when the input speech is not in optimal circumstances, for example when the user is under emotional or stress conditions. The objective of this paper is measuring the effects of stress on speech to ultimately try to mitigate its consequences on a speaker recognition task. On this paper, we develop a stress-robust speaker identification system using data selection and augmentation by means of the manipulation of the original speech utterances. An extensive experimentation has been carried out for assessing the effectiveness of the proposed techniques. First, we concluded that the best performance is always obtained when naturally stressed samples are included in the training set, and second, when these are not available, their substitution and augmentation with synthetically generated stress-like samples, improves the performance of the system.
Index Terms: speaker recognition, speaker identification, emotions, stress conditions, data augmentation, synthetic stress

1. Introduction

In recent years the interest to detect and interpret emotions in speech as well as to generate certain emotions in speech synthesis have grown in parallel. It is well-known that speech recognition systems function less efficiently when the speaker is under an emotional state, and in fact, some studies consider emotions in speech as a distortion [1].

To be able to synthesize an emotion in speech, it is necessary to analyze what are the characteristics that make it different from neutral speech. The work done about emotions in speech is very extensive, analysis are performed to study what features or combinations of them carry more information about emotions improving speech recognition rates [2], and some works aim to model emotions in speech by manipulating systematically some of the parameters of human speech, generating synthetic speech that simulates emotions [3].

Moreover, the record-keeping of databases with emotional and neutral speech is difficult as they are either recorded by actors simulating speech under those emotions, or by people under actual emotions, which could be complicated to induce. Nevertheless, stress is not considered a proper emotion, although it is intimately related to anxiety and nervousness, it is a state of mental or emotional tension resulting from adverse or demanding circumstances.

There is plenty of work about the effects of emotions in Automatic Speech Recognition (ASR) or classification of emotions in speech, but there is few work of the effects of emotions in Speaker Recognition (SR), not to mention about stressed speech on SR. Stressed speech is hard to simulate as it appears together with physical changes such as the increase of heart rate and skin perspiration. There are also hardly any databases in which stressed speech is either simulated or recorded under real conditions, along with the difficulty involved in the labelling process.

The research performed on this paper is part of a project called 'BINDI: Smart solution for Women's safety XPRIZE' by UC3M4Safety group [4]. The UC3M4Safety is a multidisciplinary team for detecting, preventing and combating violence against women from a technological point of view. The goal of this project is to develop a wearable solution that will detect a user's panic, fear and stress through physiological sensor data, speech and audio analysis and machine-learning algorithms. The ability to detect whether the voice belongs to the user or to anyone else, even under stress conditions is where this research comes in.

In this paper we want to analyze how does stress in speech affect speaker recognition rates. We aim to find techniques for strengthening speaker recognition systems, either neutralizing the effects of stress or being able to model and synthesize it from neutral speech, to create synthetically stressed speech using data augmentation techniques.

The rest of the paper is organized as follows: in Section 2 we describe the state of the art in speaker recognition and discuss features and classifiers used in literature. In Section 3, we explain the methodology followed for the feature extraction and the data augmentation techniques. Section 4 refers to the experimental set-up and results, and finally in Section 5 we discuss the conclusions and future work.

2. Speaker Recognition Related Work

Speaker Recognition is the automatic detection of a person from the characteristics of their voices (voice biometrics) [5]. We can distinguish two tasks, Speaker Identification and Verification. The first refers to the recognition of a particular user among a known number of users (a multiclass setting), and the second aims at identifying one user versus the rest (binary setting).

2.1. Features

In the literature, many features are usually used for Speaker Recognition, for example: Mel-Frequency Cepstral Coefficients (MFCC) -due to their low complexity and high performance in controlled environments-, Phonetic and Prosodic features [6] or the Linear Prediction coefficients (LP) [7]. All of these features exhibit good performance in the task when used in neutral or emotionless speech.

For speaker recognition under stress conditions, however, there is hardly any previous work, even though, MFCCs, along with Linear Frequency Cepstral Coefficients (LFCC) and Linear Prediction Cepstral Coefficients (LPCC) are cited as important features [8], together with the Pitch, Energy and Duration, which are features that seem to differ between speakers.

10.21437/IberSPEECH.2018-4

2.2. Data augmentation

Data augmentation (DA) is a commonly used strategy adopted to increase the quantity of training data. It is a key ingredient of the state of the art systems for image and speech recognition [9]. It can act as a regularizer in preventing overfitting [10] and improving performance in imbalanced class problems [11], making the whole process more robust and achieving a better performance. It is also very useful for small data sets, as it is our case, to augment the speech database and as a consequence improve accuracy [12].

2.3. Classifiers

Methods such as Gaussian Mixture Models (GMM) are generally used for speaker recognition, Support Vector Machines are widely applied as well [13], [14]. However, several studies suggest the use of Deep Learning for speaker recognition [15] and others prove the improvement in SR performance using Convolutional Neural Networks [16].

In recent years Deep Learning algorithms have skyrocketed in many scientific fields specially when using a large number of data. But for this research, we aim to keep a balance between computational complexity and accuracy, due to the constraints that our targeted device hardware imposes and the reduced amount of data originally available. Also, preliminary tests to compare GMM, SVM and Multi-Layer Perceptron (MLP) led us to chose the later, a precursor of Deep Neutral Networks, due its better performance.

3. Methodology

In this section we describe the theoretical part of the proposed system, the extraction of the features and the data augmentation techniques. The block diagram for the training stage is represented in Figure 1. The test stage is identical with the exception of the Data Augmentation block.

3.1. Feature Extraction

The acoustic features of speech extracted from audio signals should reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style).

We worked with the features extracted in the work done by Alba Mínguez [17] within BINDI for stress detection since the database employed is the same. These are the pitch, first three formants, twelve Mel-Frequency Cepstral Coefficients and the energy of the signals. The short-term features were computed every 10 ms of audio and then a temporal integration was performed over 1s length segments, calculating the mean and standard deviation, and resulting in one feature vector per 1s audio frames, which is the rate at which the accompanying heart rate measures used for labelling stress were taken.

3.2. Data augmentation

As for our device, we would hypothetically have neutral speech for the learning step and we may find stressed speech for testing. For those reasons and regarding to the low number of samples we have, we considered the generation of a synthetically stressed database performing data augmentation for the particular case of stress conditions. To be able to produce stressed speech out of neutral utterances we carried out an analysis, first listening to the audio signals and detecting what differences could be appreciated between them, and second, measuring those differences between stressed and neutral frames for the same speakers.

As a first outcome, we realized that locution speed reflects the stress of a person, we tend to pronounce more words per second and produce longer pauses when stressed. In these same conditions, there is a tendency to rise the frequency of our voices. Thus, the speed and pitch from audio signals are two variables that we aim to modify by using the SOX library [18] in order to artificially simulate speech under stress conditions.

4. Experiments

In this section we present the construction of our system in a block by block basis: we introduce the database, the labelling strategy, the preprocessing of the data, and the experiments carried out.

4.1. Corpus database

We used the so-called VOCE Corpus Database [19], a 45-speaker recordings database in neutral and stress conditions. For each of the users, speech was recorded on 3 different scenarios: *recording*, *prebaseline* and *baseline*, which were acquired respectively, in a public speaking setting where the speaker is supposed to be under stress conditions, the speaker is reading a paper 24 hours before the speech, and again reading the same paper 30 minutes but before the public speaking setting. The heart rate (HR) was also acquired every second for the three recordings.

Table 1: *Number of samples*

Samples	Neutral	Stressed	Total
Set 1	1389	3989	5378
Set 2	1716	4858	6574
Total	3105	8847	11952

However we only used 21 speakers out of the 45 due to the lack of properly recorded HR information, noisy audios or absence of recordings. We divided these 21 speakers into two sets, *Set 1* was composed of 10 speakers whose HR were coherent with the recordings in the sense that, when a speaker was reading the heart rate remained stable, but on the public speaking setting the HR rose. *Set 2* was made out of the other 11 remaining speakers. In Table 1 the number of samples per setting are specified, each sample representing 1s audio frames.

4.2. Preprocessing

For simplicity, we begin with a conversion from stereo to mono of the audio recordings, followed by a downsampling from 44100Hz to 16000Hz to reduce the computational cost of the problem without loosing too much precision. Then, a normalization of the signals in amplitude is achieved to be able to compare between them, and finally the signals go through a voice activity detector (VAD) [20] that removes silent audio frames as those don't include valuable information to our task.

4.3. Labelling

Labelling an audio signal to determine stress presence is a delicate matter since there is not a prescribed way to do so given stress is non binary and very subjective. Taking a pragmatical perspective, once more we relied on the work done by Alba Mínguez [17] where the recordings of this corpus were labeled

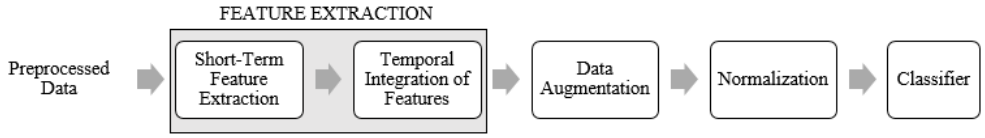

Figure 1: *Block Diagram of the system*

according to each user's heart rate (HR). Every 1s audio frame is labelled as stressed or neutral using two different heart rate thresholds. We selected the binarization threshold that gave better results in their report, which was the 75% percentile of the HR of the user.

4.4. Balancing the data

Soon we realized that the data instances were not balanced for each speaker. An adjustment needs to be made for each set and conditions to get consistent estimates as all classes have the same importance. Nevertheless, the use of an over-sampling technique would have a big drawback in our case because some users have significantly more samples than others, and this would create too many artificial samples. To cope with this problem we cropped randomly the neutral samples by a threshold of 120 samples for both sets, and stressed samples over a threshold of 300 samples. Applying an over-sampling technique (in particular, SMOTE) [11] to the new cropped data culminated in new samples resulting in a balanced data set.

4.5. Preliminary Experiments

Originally, for an initial experimental set-up we used the data available for Sets 1 and 2 (21 speakers). This preliminary experiment is made to observe the behaviour of mismatch conditions' experiments on the speaker recognition rate. First of all, we divided the data in neutral (NS) and stressed speech (S) and experimented training with one type of speech and testing with the other, and then mixing both types. In order to get reliable results, these experiments were repeated 50 times where, in each repetition, at least 50% data was randomly chosen for testing. The results in terms of accuracy (percentage of audio segments correctly classified) are in Table 2.

Table 2: *Results for match and mismatch settings*

Train	Test	Mean (%)	Std (%)
NS	NS	96.73	0.33
	S	79.21	0.90
S	S	95.87	0.28
	NS	90.89	0.49
MIX	MIX	96.05	0.12

As a first conclusion, match settings are beneficial and mismatch are not, as was expected. When training with neutral and testing with stress, accuracy decreases, so it seems that stressed speech does have different characteristics compared to neutral speech. On the contrary, when training with stress and testing with neutral utterances, the decrease in accuracy is not that important, leading us to think that stressed speech could be sparse data in which neutral speech could be contained but not vice versa. About the mixed conditions experiments, the accuracy reached a 96.05%, achieving a good result for this particular task.

4.6. Synthetic Stress

We performed an analysis to measure the differences between the mean pitch from neutral to stressed audio frames for each speaker using VOICEBOX [20], and we also estimated the average elocution speed for each user. To do this, we obtained an automatic transcription of each of the recordings by using Google Speech Recognition [21] and computed afterwards the mean number of words per second.

The differences of pitch from neutral to stressed speech were between -2% and +7%, increasing an average of 2.2%. As regards to the elocution speed, subjectively, it seems to increase in stressed speech, but our analysis gave us the opposite conclusion. The number of words per second was higher when the user was reading a text, 2.2 words/s in mean, than when the speaker was performing an oral presentation, 1.85 words/s. By listening to the signals, we determined that the words were pronounced faster but there were more pauses in between them, leading to a lower elocution rate in overall.

Thus, we have changed the locution speed and the pitch from the original database, to produce synthetically stressed samples of speech. The pitch was modified in steps of [-6%, -3%, +3%, +6%], and the signals were reproduced at the following speeds [-20%, -15%, -10%, -5%]. All these modifications are applied to the original sets and result in an augmentation of data, one new synthetic set per modification.

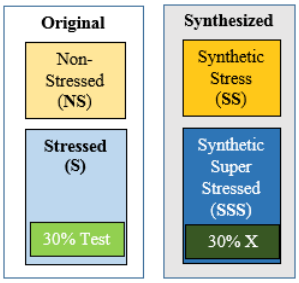

Figure 2: *Original and Synthetic databases scheme*

Figure 2 presents an schematic of the original data and its counterpart synthesized one. SS and SSS represent the synthetically stressed collection from NS and S respectively. For the experimental set-up we always use the same test set, a 30% of the samples of original stressed speech, which is represented in green. Additionally, the same 30% in the synthetic super-stressed set was removed to achieve a more accurate comparison between experiments. In Figure 3 we enumerate the data used for training for each of the experiments.

The results of the pitch modifications experiments for Set 1 are presented in Figure 4, and the ones for speed modifications for the same set in Figure 5. From these figures, we can see that the changes that improve the accuracy the most are Pitch -3% and +3%, and although in speed the results are very similar, the one that in general works worse is Speed -20%.

From the experiments carried out for Set 1, we decided to perform the following modifications to Set 2 pitch [-3%, +3%] and signal speed [-15%, -10%, -5%].

Case	1	2	3	4	5	6	7	8	9	10
Training data	NS	S(70%)	NS + S(70%)	NS + SS	NS + SSS(70%)	SS	SSS(70%)	NS + S(70%) + SSS(70%)	NS + S(70%) + SS + SSS(70%)	NS + SS + SSS(70%)

Figure 3: *Equivalence between training data and number of experiment.*

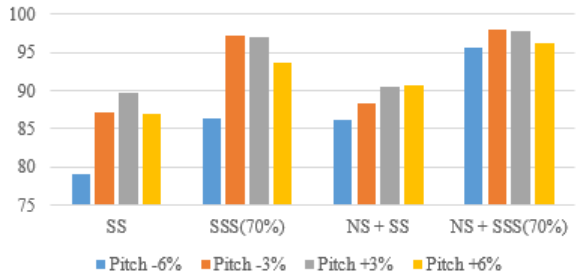

Figure 4: *Results for Synthetic datasets, Set 1 Pitch Modifications*

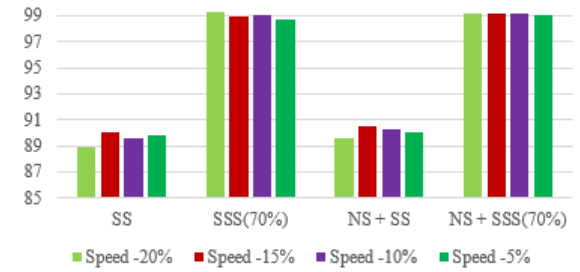

Figure 5: *Results for Synthetic datasets, Set 1 Speed Modifications*

After that, we joined Sets 1 and 2, transforming the problem in a 21-speaker SR task, and combined all the synthetic stress data, multiplying by 5 the original dataset. We repeated the 10 experiments 20 times in order to obtain reliable results. The outcome is available in Table 3, and the equivalence between Training Set and Case number is shown in Figure 3.

Table 3: *Results for extensive experimentation*

Case	Set 1 Mean	Set 1 Std	Set 1+2 Mean	Set 1+2 Std
1	89.71	0.56	78.55	0.6
2	98.59	0.16	97.37	0.21
3	98.48	0.23	97.21	0.26
4	89.97	0.39	80.46	0.53
5	99.93	0.05	99.16	0.11
6	89.72	0.53	78.19	0.71
7	99.88	0.07	99.21	0.13
8	99.91	0.07	99.45	0.08
9	99.94	0.06	99.22	0.11
10	99.91	0.07	98.97	0.14

There are two types of experiments in Table 3: those where we substitute data and the ones where we augment data. As for substituting the original set by a synthetically stressed one, we have experiments 6 and 7, to be compared with experiments 1 and 2 respectively. Data substitution achieves similar results to the experiments with original data when using synthetic data obtained from NS speech for training (case 1 vs. case 6) and better recognition rates when using synthetic data obtained from S speech for training (case 2 vs. case 7).

The data augmentation experiments are 3, 4, 5, 8, 9 and 10. The outcome is indeed positive, the best results are achieved in experiment 8 with a 99.45% of accuracy for Sets 1 + 2. These results show us that augmenting the data boosts the SR rate.

One of our objectives with these experiments was that experiment 4 could outperform experiment 2, meaning that we accomplished the task of generating appropriate synthetically stressed speech out of neutral. That goal has not been achieved, but at least in Table 3 experiment 4 is better than 6 which, in turn outperforms 1, for Set 1 and for Set 1+2, settling that stressing speech synthetically and using it as training data alongside with the original data, increases the performance of the SR system.

5. Conclusions and future work

In this research our goal was to analyze how stressed speech affects Speaker Recognition systems. We have identified a problem, which is that stressed speech affects negatively when SR systems are trained only with neutral speech.

In the experiments for data substitution, depending on the difference between the synthetic data and the original one, the substitution outperforms the original data. Besides, the modifications over the speed of the audio signals work better for substituting audio utterances than the modifications in pitch.

As regards to the experiments for augmenting the database with artificial stress, we can conclude that the generation of different synthetically stressed utterances of speech and its addition to the database improves substantially the SR results, reaching a 99.45% of accuracy rate in experiment 8 for Sets 1+2.

Due to limited time and computational power, several experiments and methods remained unexplored and left for future work:

- Our target in this research is a Speaker Identification task, a multiclass problem. However, the objective of the device to be built in BINDI is a Speaker Verification system. These two approaches are not straightforwardly comparable but we believe that the problems and solutions can be translated to one another.
- To simulate a real environment in which the recorded voice is not clean, we could add noise to the same database used and analyze its effect.
- Further analyzing the differences between neutral and stressed speech could lead us to finding new modifications to perform to neutral speech to transform it into an appropriate synthetically stressed speech.
- Implementing new methods for recording stressed speech using BINDI such as Stroop Effect games [22] in which the speaker should experiment stress.
- With the use of data augmentation techniques we have collected a much larger database and we could therefore employ more powerful Deep Learning algorithms in the future.

6. Acknowledgements

This work is partially supported by the Spanish Government-MinECo projects TEC2014-53390-P and TEC2017-84395-P.

7. References

[1] J. H. Hansen and S. Patil, "Speaker classification," C. Müller, Ed. Berlin, Heidelberg: Springer-Verlag, 2007, ch. Speech Under Stress: Analysis, Modeling and Recognition, pp. 108–137. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74200-5_6

[2] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," vol. 3, 12 1996.

[3] I. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," vol. 93, pp. 1097–108, 03 1993.

[4] "UC3M4SAFETY - Multidisciplinary team for detecting, preventing and combating violence against women," 2017. [Online]. Available: http://portal.uc3m.es/portal/page/portal/inst_estudios_genero/proyectos/UC3M4Safety

[5] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2018.

[6] E. Shriberg, *Higher-Level Features in Speaker Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 241–259. [Online]. Available: https://doi.org/10.1007/978-3-540-74200-5_14

[7] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep 1997.

[8] D. A. Reynolds, T. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models." vol. 10, no. 1, p. 19–41, 2000.

[9] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," 2013. [Online]. Available: http://www.cs.toronto.edu/~ndjaitly/jaitly-icml13.pdf

[10] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *ICDAR*, 2003.

[11] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011. [Online]. Available: http://arxiv.org/abs/1106.1813

[12] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Computer Science*, vol. 112, pp. 316 – 322, 2017.

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[14] K. A. Abdalmalak and A. Gallardo-Antolín, "Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers," *Neural Computing and Applications*, vol. 29, no. 3, pp. 637–651, Feb 2018.

[15] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 194–205, Jan 1994.

[16] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," pp. 686–690, 01 2014.

[17] A. Mínguez-Sánchez, "Detección de estrés en señales de voz *[Stress detection in voiced signals]*," p. 86, 06 2017. [Online]. Available: https://github.com/minguezalba/Stress_Detection

[18] R. Bittner, E. Humphrey, and J. Bello, *PySOX: Leveraging the Audio Signal Processing Power of SOX in Python*. International Conference on Music Information Retrieval (ISMIR-16), 8 2016.

[19] A. Aguiar, M. Kaiseler, C. M, J. Silva, M. H, and P. Almeida, "Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments." 05 2014. [Online]. Available: https://repositorio-aberto.up.pt/bitstream/10216/85669/2/133351.pdf

[20] M. Brookes, "Voicebox: Speech processing toolbox for matlab [software]," 01 2011. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[21] A. Zhang, "Speech recognition (version 3.8) [software]," 2017. [Online]. Available: "https://github.com/Uberi/speech_recognition"

[22] J. Ridley Stroop, "Studies of interference in serial verbal reactions," in *Journal of Experimental Psychology: General*, vol. 121, 03 1992, pp. 15–23.