INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Depression Detection Using Automatic Transcriptions of De-Identified Speech

Paula Lopez-Otero[1], *Laura Docio-Fernandez*[1], *Alberto Abad*[2,3], *Carmen Garcia-Mateo*[1]

[1]AtlantTIC Research Center, Multimedia Technologies Group, University of Vigo, Spain
[2]L[2]F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[3]IST - Instituto Superior Técnico, University of Lisbon, Portugal

[1] {plopez,ldocio,carmen}@gts.uvigo.es
[2,3] alberto.abad@l2f.inesc-id.pt

Abstract

Depression is a mood disorder that is usually addressed by outpatient treatments in order to favour patient's inclusion in society. This leads to a need for novel automatic tools exploiting speech processing approaches that can help to monitor the emotional state of patients via telephone or the Internet. However, the transmission, processing and subsequent storage of such sensitive data raises several privacy concerns. Speech de-identification can be used to protect the patients' identity. Nevertheless, these techniques modify the speech signal, eventually affecting the performance of depression detection approaches based on either speech characteristics or automatic transcriptions. This paper presents a study on the influence of speech de-identification when using transcription-based approaches for depression detection. To this effect, a system based on the global vectors method for natural language processing is proposed. In contrast to previous works, two main sources of nuisance have been considered: the de-identification process itself and the transcription errors introduced by the automatic recognition of the patients' speech. Experimental validation on the DAIC-WOZ corpus reveals very promising results, obtaining only a slight performance degradation with respect to the use of manual transcriptions.

Index Terms: depression detection, speech de-identification, global vectors

1. Introduction

Depression is a common mental disorder that causes people to experience depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or low self-worth, disturbed sleep or appetite, and poor concentration [1]. Nowadays, this disorder is usually addressed by outpatient treatments in order to favour patient's inclusion in the society in a cost-effective manner [2]. These treatments are usually accompanied by an intensive supervision carried out by practitioners, which is highly time demanding. This leads to a need for approaches that can help to monitor the emotional state of patients via telephone or the Internet. The use of speech processing techniques for this purpose is suitable, since speech is a biomarker of depression that can be measured remotely and non-invasively [3]. One of the drawbacks of such techniques is the privacy concern raised by the transmission of speech and its further storage in servers. This issue can be overcome by means of de-identification, which is a process by which a data custodian alters or removes identifying information from a dataset, making it harder for users of the data to determine the identities of the data subjects [4]. The most extended technique for speaker de-identification consists in applying voice conversion techniques [5, 6, 7, 8] in order to modify the voice characteristics of a speaker in a way that, afterwards, they sound like a different speaker.

The use of speaker de-identification solves the aforementioned privacy issue, but the influence of de-identification in the perception of diseases or mental disorders that affect speech production, such as depression, has not been addressed in depth yet. This procedure would be very time-consuming when this evaluation is performed by specialists. Fortunately, there is a considerable amount of work in the field of automatic depression detection using speech. Given that speech carries information both in the message and in its acoustic characteristics, the literature covers depression detection approaches based on acoustic and prosodic characteristics extracted from the speech signal [3, 9, 10, 11], and also based on natural language processing techniques applied to the text transcription [12, 13, 14].

This paper presents a study on how speaker de-identification affects the performance of a depression detection system based on speech transcriptions. The interest of this analysis is twofold: it addresses the effect of de-identification in automatic speech recognition (ASR) systems, and also the suitability of performing depression detection using automatic transcriptions with errors.

In order to carry out the proposed investigation, a depression detection approach based on global vectors (GloVe) [15] is presented in this paper. This word embedding technique allows the representation of words in a high-dimensional space where the similarity in meaning between two words is directly related to the distance of their corresponding vectors. Additionally to this technique, already used in [13] for depression detection, a word weighting strategy is proposed in order to mitigate the effect of errors in the automatic transcriptions. This weighting gives more relevance to those words that are less frequent in the training corpus at the expense of more common terms such as function words [16].

Experiments were performed in the framework of the Audio/Visual Emotion Challenge (AVEC) 2016 [17]. The experimental validation showed that the proposed approach for depression detection using GloVe embedding achieves very competitive results when using manual transcriptions. In addition, a comparison between these results and those obtained using automatic transcriptions, either de-identified or not, showed a slight degradation produced by these procedures, but the results are very encouraging.

The rest of this paper is organised as follows: Section 2 provides an overview of the experimental protocol followed in this paper; Section 3 presents the proposed approach for depression detection using transcriptions; the experimental framework used in this paper is described in Section 4; Section 5 describes the experiments and results; and Section 6 ends with some conclusions and future work remarks.

http://dx.doi.org/10.21437/Interspeech.2017-1201

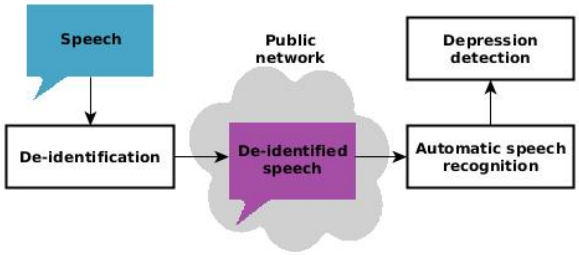

Figure 1: *Block diagram of the experimental protocol.*

2. System overview

Figure 1 presents an overview of the proposed strategy for depression detection on de-identified speech. First, speech is de-identified so that the privacy of the speaker is preserved for its transmission in a public network. This is achieved using a voice conversion approach based on frequency warping (FW) plus amplitude scaling (AS). Then, speech is transcribed using an ASR approach employing long short term memory (LSTM) acoustic models, since this type of recurrent neural network (RNN) architecture has shown to achieve great improvements in recognition performance [18]. Finally, the depressive state of the speaker is obtained from the transcribed speech.

The rest of this Section provides an overview of the speaker de-identification and ASR approaches considered in this paper. Since the depression detection approach deserves a more detailed explanation, it is fully described in the next Section.

2.1. Speaker de-identification approach

As briefly mentioned previously, voice conversion was used in this work to achieve speaker de-identification. It consists in modifying the voice characteristics of a source speaker in order to make it sound like a different target speaker. Typical techniques for voice conversion require a parallel corpus between the source and target speakers to train the transformation function, which is not always available. Thus, the technique proposed in [7] is used in this work, since it does not has that restriction. This approach applies a linear transform in the cepstral domain based on FW combined with AS:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

where $\mathbf{x}$ is a Mel-cepstral vector, $\mathbf{A}$ represents a FW matrix, $\mathbf{b}$ is an AS vector, and $\mathbf{y}$ is the transformed version of $\mathbf{x}$.

A training stage is typically performed in FW+AS strategies in order to obtain the parameters of $\mathbf{A}$ and $\mathbf{b}$. In the method proposed in [7], instead of training these parameters, they are manually defined following some guidelines. First, the FW curve is simplified and defined piecewise using three linear functions as shown in Figure 2: the discontinuities of the curve are set at frequencies f_a and f_b; α is the angle between the 45-degree line and the first linear function; and β is the angle between the 45-degree line and the second linear function, defined as $\beta = k\alpha$ $(0 < k < 1)$. When α is greater (less) than 0, formants are moved to higher (lower) frequencies, resulting in a male-to-female (female-to-male) transformation function.

The AS vector $\mathbf{b}$ is defined by randomly giving values to a set of weighted Hanning-like bands equally spaced in the Mel-frequency scale [19] as fully described in [7]. Finally, FW+AS is complemented with a scaling of the fundamental frequency proportional to the value of α, since it showed to dramatically improve de-identification performance [7].

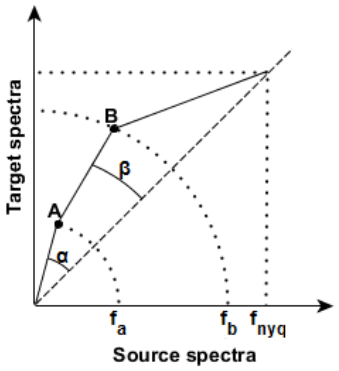

Figure 2: *Piecewise linear approximation of a FW function.*

2.2. Automatic speech recognition approach

A large vocabulary continuous speech recognition (LVCSR) system has been built using the Kaldi speech recognition toolkit [20]. Specifically, the standard Kaldi LibriSpeech recipe was used to build a set of hybrid Long Short Term Memory (LSTM) / Hidden Markov Model (HMM) acoustic models. In this work, a LSTM architecture with 3 LSTM hidden layers followed by two fully-connected feed-forward layers was used. Each LSTM layer has 1024 memory cells and 256 output units in projection. The acoustic input frames to the LSTM consist in 40 dimensional Mel frequency cepstral coefficients (MFCCs) plus 100 dimensional i-vectors [21], and the input to the LSTM is a single acoustic frame with a context splicing of 5 frames, and the output state label of the LSTM is delayed by 5 frames. These acoustic models were trained using the LibriSpeech corpus [22], which is a read speech dataset derived from English LibriVox audiobooks[1]. The train subset consists of 280k utterances, representing approximately 960 hours of speech.

The language model (LM) used by the LVCSR is a modified Kneser-Ney smoothed 3-gram model trained using the SRILM toolkit [23] on a text corpus selected from public domain books of Project Gutenberg[2]. No pruning of the LM was done since it empirically showed better performance.

The lexicon contains the 200k most frequent words in the LibriSpeech corpus, whose pronunciations were either obtained from the CMU pronunciation dictionary (when available) or generated using the Sequitur G2P toolkit [24].

3. Depression detection from transcribed speech

The proposed approach for depression detection relies on a natural language approach for word representation named global vectors (GloVe) [15]. First, an overview of this method is given, and then the strategy for depression detection is defined.

3.1. GloVe for word representation

GloVe is a word embedding technique that learns the meaning of words in an unsupervised manner [15]. The goal of this embedding is obtaining a vectorial representation of words such that the closer the vectors, the more similar the meaning of their corresponding words. Hence, given a training corpus, i.e. large amounts of text, first a co-occurrence matrix is computed, since it is assumed that words that co-occur very often have a certain

[1] www.openslr.org/12/
[2] www.gutenberg.org

similarity in meaning. This matrix is used to train a neural network that returns a vector for each word in the training corpus.

Once a vectorial representation for the training words is obtained, given two words, it is possible to obtain their similarity in meaning by computing the distance between their corresponding vectors. Results presented in [15] show that this technique outperforms other similar approaches in different tasks.

3.2. Depression detection using GloVe embedding

As described above, GloVe representation imply that closer words have similar meaning. For depression detection, it is necessary to know which words are related to depression in order to learn a classifier that discriminates depressive states.

Given a set of training transcriptions representing the classes "depressed" and "not-depressed", first these transcriptions are split in speaker turns. Hence, let $\mathrm{T} = \{\mathrm{T}_1, \ldots, \mathrm{T}_{\mathrm{n_T}}\}$ represent the set of $\mathrm{n_T}$ training turns, where each turn T_i has an associated class $\mathrm{C}_i \in \{\text{depressed, not depressed}\}$. A turn is composed of a set of $\mathrm{n_{T_i}}$ words $\mathrm{T}_i = \{\mathrm{w}_1, \ldots, \mathrm{w}_{\mathrm{n_{T_i}}}\}$, and each of these words can be represented using its corresponding GloVe embedding $\mathrm{w}_i \sim \mathbf{w}_i$. A speaker turn can then be represented by a single vector

$$\mathbf{T}_i = \frac{1}{\mathrm{n_{T_i}}} \sum_{j=1}^{\mathrm{n_{T_i}}} \frac{\mathbf{w}_j}{|\mathbf{w}_j|}. \tag{2}$$

Eq. (2), however, gives the same importance to all the words in a turn. Let us assume that the most frequent words (which are, in general, function words) in the training corpus are less relevant than those that are less frequent. Hence, a weight α can be assigned to a word such that:

$$\alpha_i = \frac{1}{\#\mathrm{w}_i} \tag{3}$$

where $\#\mathrm{w}_i$ represents the number of occurrences of the word w_i in the training corpus. Finally, replacing Eq. (2), the vector of a speaker turn is computed as

$$\mathbf{T}_i = \frac{1}{\mathrm{n_{T_i}}} \frac{1}{\sum_{j=1}^{\mathrm{n_{T_i}}} \alpha_j} \sum_{j=1}^{\mathrm{n_{T_i}}} \alpha_j \frac{\mathbf{w}_j}{|\mathbf{w}_j|}. \tag{4}$$

It must be noted that, in Eq. (4), the word weights are normalised so that they sum 1.

As suggested in [13], the vectors $\mathbf{T}$ can be transformed in order to obtain a representation of the speech turns in a low dimensional subspace $\mathbf{T}'$, using techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) or zero-phase component analysis (ZCA) whitening.

After this procedure, the training corpus is composed by a set of dimensionality reduced vectors $\mathbf{T}' = \{\mathbf{T}'_i, \ldots, \mathbf{T}'_{\mathrm{n_T}}\}$ with its corresponding set of depressed/not-depressed labels C as defined above. These data vectors can be used to train a classifier that discriminates between the two classes. Thus, in the proposed system, a support vector machine (SVM) classifier has been trained using the vectors $\mathbf{T}'$.

Then, for each turn vector $\mathbf{T}'_i$, the SVM is used to obtain the score s_i corresponding to the distance from the vector to the hyperplane defined by the SVM. Hence, a full speech transcription is represented by the set of scores corresponding to each of its speaker turns. In order to obtain a single score for each transcription, the median of their corresponding speaker turn scores is computed. After this, the transcription scores are further processed by a linear logistic regression function in order to obtain well-calibrated scores [25]. This calibration stage is trained with the Bosaris toolkit [26] using the transcription scores of the training set. Given that this procedure is expected to produce well-calibrated scores, the theoretical optimum Bayes threshold can then be used for making hard decisions.

During the test phase, given a test transcription, its estimated depressed/not-depressed label can be obtained by obtaining the vectors for each speaker turn, computing the median of their scores given by the SVM, applying the calibration parameters, and comparing with the decision threshold.

4. Experimental framework

Depression detection experiments were performed using the DAIC-WOZ corpus, which compiles a series of recordings of clinical interviews in English language designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post traumatic stress disorder [27]. These documents were manually transcribed, and a label depressed/not-depressed was assigned to each session.

Table 1: *Summary of DAIC-WOZ database: total number of sessions, sessions of classes depressed (D) and not-depressed (ND), and duration of the recordings.*

Partition	Sessions (D/ND)	Duration
Train	107 (21/86)	26 h 54 min
Development	35 (7/28)	10 h 1 min

Since this database was used in the Audio/Visual Emotion Challenge (AVEC) 2016 for depression detection [17], the experimental protocol designed for that evaluation was used. The data used for these experiments is summarised in Table 1. Since the depression labels of the test data were not provided with the database, the test sessions were not used in these experiments, so performance was assessed on the development sessions. The evaluation measures used in these experiments were the classification accuracy and the mean F-score, defined as the mean of the F-score of classes depressed and not-depressed [17].

5. Experimental results

Before describing the experiments, some configuration aspects must be established. First of all, a preprocessing of the transcriptions was done. Since the interviews included in DAIC-WOZ were conducted by an interactive agent and a participant, both sides of the conversation are included in the manual transcriptions. As shown in [13] and [14], the interviewer's side of the conversation provides valuable information about the depression state of the participant. For example, when the participant answers "yes" to the question "Have you been diagnosed with depression?" the follow-up question "How long ago were you diagnosed?" is asked. Using this information to build a depression detection system would over-fit the task proposed in AVEC 2016 challenge so, in all the experiments reported in this paper, the interviewer's side of the conversation was discarded. For the participant's side of the conversation, the speaker turns are defined as all the words spoken without being interrupted by the interviewer. In addition, the interviews include some semantic information such as *laughter*, *sigh* or *deep breath* that was removed from the transcriptions in order to enable a fair comparison with the automatically transcribed data, since these semantic cues would not be available in that case.

With respect to the de-identification strategy, the parameters $\mathrm{f_a}$, $\mathrm{f_b}$ and k were set to 700 Hz, 3000 Hz and 0.5 according

to [7], while the parameter α was set to $\pi/24$ for male-to-female conversion and $-\pi/24$ for female-to-male conversion. This configuration exhibited a de-identification accuracy of 88.1% in the experiments reported in [7]. In the depression detection strategy, a pre-existing GloVe model of dimension 50 was used, which was trained using a Twitter corpus of 2 billion tweets with a vocabulary of 1.2 million words[3].

Once the aforementioned design decisions were established, a set of experiments was defined:

- Wizard of Oz (WoZ) experiment. Given the manual transcriptions as provided with the database, depression classification experiments were performed as described in Sec. 3. Different representation techniques were assessed, namely the original vectors obtained using Eq. (4), PCA, LDA and ZCA. The top-performing representation was used in the rest of the experiments, and its performance was used to establish a comparison with the rest of the experiments and with other results found in the literature. An experiment WoZnoW (no weighting) was also performed to evaluate the goodness of the word weighting strategy defined in Eq. (4).
- ASR experiment. The ASR system described in Section 2.2 was used to obtain automatic transcriptions of the development recordings. This experiment allows to quantify the performance loss caused by the use of automatic transcriptions instead of manual annotations. Once the recordings were transcribed, the depression classification approach trained with the WoZ data was used to obtain depression labels for the automatic transcriptions.
- De-ID+ASR experiment. First, the development transcriptions were de-identified as described in Sec. 2.1. After that, the same procedure as in the ASR experiment was performed. The performance loss caused by the de-identification procedure can be quantified by comparing the results of this experiment and the ASR one.

Regarding the WoZ experiment, Table 2 shows that depression detection improves when using the word weighting approach proposed in Eq. (4). The best results were achieved when representing the vectors in a PCA subspace of dimension 40. It can also be noted that these results are superior to those presented in [13] when using GloVe and PCA on the participant's side of the conversation only, where a mean F-score by 62% was reported. This proves the validity of the proposed approach for depression detection using speech transcriptions.

Table 2 shows that the WER achieved on the ASR experiment was rather high. Nevertheless, the degradation in depression detection performance was slight since, compared to the WoZ experiment, only an additional transcription was misclassified. This suggests that the vector representations are considerably robust to transcription inaccuracies.

Lastly, the De-ID+ASR experiment was conducted. Table 2 shows that the WER increased by 4% compared to that achieved using the original data, which reveals a slight degradation on ASR results caused by de-identification.In addition, depression detection experiments resulted in an additional misclassification when compared with the ASR experiment, that is, two more errors than in the WoZ experiment, suggesting that the effect of de-identification followed by automatic transcriptions in this task is quite limited.

[3] https://nlp.stanford.edu/projects/glove/

Table 2: *Results achieved on the experiments in terms of accuracy, mean F-score and F-score of classes depressed (D) and not-depressed (ND). WER is shown when applicable.*

Experiment	WER	Accuracy	F-score: Mean (D/ND)
WoZ	-	85.7% (30/35)	73.0% (54.5%/91.5%)
WoZnoW	-	80.0% (28/35)	66.9% (46.2%/87.7%)
ASR	32.9%	82.9% (29/35)	65.0% (40.0%/90.0%)
De-ID+ASR	37.3%	80.0% (28/35)	62.2% (36.4%/88.1%)

6. Conclusions and future work

The impact of speaker de-identification in a depression detection system based on automatic transcriptions was assessed in this paper. An experimental protocol was defined starting from that defined for AVEC 2016 challenge, which allowed to find out the influence of the different sources of error involved in the process, namely the de-identification process and the errors in automatic transcriptions.

A strategy for depression detection was presented in this paper, which was based on GloVe embedding for word representation. Differently to other systems proposed in the literature, the system described in this work gives more or less relevance to each word according to its frequency (less common words are assumed to be more relevant). Experimental validation showed an improved performance of the system when weighting the vectors in the proposed manner. This system outperforms other approaches found in the literature.

The proposed depression detection system was used to perform depression detection using automatic transcriptions of original and de-identified speech. The results achieved were very encouraging as, despite the high WER observed on the transcriptions, depression detection results just suffered a slight degradation. In future work, an in-depth analysis of the automatic transcriptions will be done in order to discover which errors affect the most the performance of text-based depression detection approaches. For this purpose, simulating transcriptions with different WERs would be interesting. The use of automatic transcriptions for training the depression detection approach in an unsupervised manner will also be assessed.

Future work also includes the assessment of the impact of de-identification on depression detection systems based on acoustic and prosodic characteristics. In addition, combining the depression detection system proposed in this paper with others using acoustic and/or prosodic information will be investigated, since this multimodal approach might lead to a boost in performance.

7. Acknowledgements

This research was funded by the Spanish Government (project 'TraceThem' TEC2015-65345-P), the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC' CN2012/160, the Portuguese Government through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, the European Regional Development Fund (ERDF) and the COST Action IC1206. The authors would like to thank Carmen Magariños for providing the speech de-identification software.

8. References

[1] M. Marcus, M. Yasamy, M. van Ommeren, D. Chisholm, and S. Saxena, "Depression: A global public health concern," World Health Organization, Tech. Rep., 2012.

[2] J. Bedell, R. Hunger, and P. Corrigan, "Current approaches to assessment and treatment of persons with serious mental illness," *Professional Psychology: Research and Practice*, vol. 28, no. 3, pp. 217–228, 1997.

[3] N. Cummins, "Automatic assessment of depression from speech: Paralinguistic analysis, modelling and machine learning," Ph.D. dissertation, The University of New South Wales, 2016.

[4] S. Garfinkel, "De-identification of personally identifiable information," National institute of standards and Technology (NIST), U.S. Department of Commerce, Tech. Rep., 2015.

[5] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *IEEE workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 529–533.

[6] M. Abou-Zleikha, Z.-H. Tan, M. Christensen, and S. Jensen, "A discriminative approach for speaker selection in speaker de-identification systems," in *Proceedings of 23rd European signal processing conference (EUSIPCO)*, 2015, pp. 2147–2151.

[7] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, D. Erro, E. Banga, and C. Garcia-Mateo, "Piecewise linear definition of transformation functions for speaker de-identification," in *Proceedings of First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016, pp. 1–5.

[8] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E. Banga, D. Erro, and C. Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," *Computer Speech and Language*, 2017.

[9] J. Williamson, T. Quatieri, B. Helfer, G. Ciccarelli, and D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of AVEC'14*, 2014.

[10] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," in *Proceedings of IWBF*, 2014, pp. 1–6.

[11] ——, "A study of acoustic features for the classification of depressed speech," in *Proceedings of MIPRO*, 2014, pp. 1331–1335.

[12] J. Correia, I. Trancoso, and B. Raj, "Detecting psychological distress in adults throgh transcriptions of clinical interviews," *Lecture Notes on Artificial Intelligence*, vol. 10077, pp. 162–171, 2016.

[13] J. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 3–10.

[14] L. Yang, D. Jiang, L. He, E. Pei, M. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 89–96.

[15] J. Pennington, R. Socher, and C. Manning, "GloVe: global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[16] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 3–10.

[18] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of Interspeech*, 2014, pp. 338–342.

[19] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernáez, "Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations," *Computer Speech and Language*, vol. 30, pp. 3–15, 2015.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[21] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 5206–5210.

[23] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[24] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, 2008.

[25] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[26] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," Tech. Rep., 2011. [Online]. Available: https://sites.google.com/site/nikobrummer

[27] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, and D. Traum, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of 9th Language Resources and Evaluation Conference (LREC)*, 2014, pp. 3123–3128.