Noise-Robust Detection of Whispering in Telephone Calls Using Deep Neural Networks

Aleksandr Diment, Mikko Parviainen, Tuomas Virtanen

Tampere University of Technology
Department of Signal Processing
Korkeakoulunkatu 1, 33720, Tampere, Finland
firstname.lastname@tut.fi

Roman Zelov, Alex Glasman

Behavox
Level39, One Canada Square, London E14 5AB
firstname.lastname@behavox.com

***Abstract*—Detection of whispered speech in the presence of high levels of background noise has applications in fraudulent behaviour recognition. For instance, it can serve as an indicator of possible insider trading. We propose a deep neural network (DNN) -based whispering detection system, which operates on both magnitude and phase features, including the group delay feature from all-pole models (APGD). We show that the APGD feature outperforms the conventional ones. Trained and evaluated on the collected diverse dataset of whispered and normal speech with emulated phone line distortions and significant amounts of added background noise, the proposed system performs with accuracies as high as 91.8%.**

I. Introduction

Whispering is a form of speech, whose production does not involve vibration of vocal cords. As opposed to normal speech with harmonic excitation, whispering is produced with a broad-band noise. Due to the differences in their production, normal and whispered speech signals differ noticeably. However, in the presence of background noise the signal-to-noise ratio of whispered speech can be so low that its automatic detection becomes challenging.

A good example of an industrial application of whispering detection is within a system that provides indicators of potential insider trading. Many enterprises are concerned about their information security and therefore are interested in detecting its breaches. Insider trading is just one of the forms of such breaches, and it refers to the misuse and trading of confidential information about the company by the individuals with access to that information. Such activity is illegal in many countries.

Whispering during a phone call on the trading floor of a capital markets business is unusual behaviour that can serve as an indicator of possible insider trading activity. Companies providing solutions for identifying these possible security breaches (e.g. Behavox, `http://behavox.com`) are interested in detecting segments of a phone call conversation that contain whispering in order to incorporate this information into a broad multi-modal social connection analysis system.

Whispering detection has other applications as well. For instance, lifelogging solutions can incorporate audio-based user activity detection that could distinguish among others speaking and whispering in order to capture connotations of speech [1]. Detecting whether the speech is produced normally or by whispering is also useful in automatic speech recognition [2], where different models can be selected based on the output of the predictor.

Previously, methods for detecting whispered speech in the presence of normal phonation were proposed for noise-free signals with use of unsupervised techniques operating on energy and periodicity features [3]. Another method [2] employs GMM-based detection of vocal effort (including whispering, soft, normal and loud) trained on MFCCs and spectral tilt features in presence of rather insignificant amount of noise (SNR 20 dB) in order to facilitate model selection for speech recognition. With activity detection application in mind, Yatani and Truong [1] have addressed recognition of whispering and other modes of vocalisation along with other human activities observed from clean noiseless audio with a set of temporal, spectral and cepstral features and Support Vector Machines.

In this paper, we focus whispering detection in realistic conditions with significant amount of background noise — a problem not addressed previously. We propose to perform the detection whispered speech recorded over a phone line in an office environment by means of deep neural networks (DNN) trained on the phase- and magnitude-based features, shown to be applicable for speech processing. We perform the detection of whispering against both normal speech and pauses filled with only the background noise.

The paper is organised as follows. The implementation details of the DNN-based classifier are outlined in In Section II, along with the proposed set of features motivated by previous studies in whispered speech production. Thereupon, in Section III we present the whispered vs. non-whispered speech corpus collected for this work. The system is thereupon evaluated in Section IV, followed by conclusions in Section V.

II. Methodology

A. Problem definition

We address the problem of detecting whispered speech segments from a continuous audio recorder over a phone line in a realistic office scenario. The audio signal contains both normal and whispered speech, as well as noise-filled pauses when the person on the phone was not producing any speech. The office environment introduces significant levels of

added background noise, and the recording is corrupted with distortions as a result of compression in the cellular phone channel.

We formulate the whispering detection as a classification problem, where acoustic features are extracted from short audio segments, and each segment is classified to either whispering or non whispering using a supervised classifier. The task is to develop a system that detects segments of whispered speech from the recording and outputs their timing. The system is developed to include state-of-the-art features and machine learning methodology.

B. Features

Due to the noisy and distorted nature of the realistic data and dynamic range variations, a simple energy thresholding or similar techniques are not applicable for whispering detection, as became apparent in the initial experiments. Instead, a set of spectral features, which incorporate both the magnitude and the phase, shown to be important for speech analysis, is proposed based on previous studies and preliminary experiments:

- **all-pole group delay** [4]: 15 APGD coefficients with first derivatives, LPC order 18. The main aspect of the method is to calculate the group delay function from all-pole models of a signal, formed by linear prediction. The method has been used in formant extraction and speaker recognition [4]. Recently, the feature has been shown to work well within the DNN framework [5]. Changes in vocal effort have been shown to affect the first formant of the speech signal [6], and the APGD feature is well suited for describing formant variations. Studies of whispered speech production [7], [8], [9] have made observations on upward shift in the resonance frequencies of whispered vowels, lower energy in linguistically voiced consonants, and greater spectral flatness in whispered speech compared with normal speech. Motivated both by the differences in spectrum of whispered and normal speech in general and the importance of incorporating phase part of the spectrum into speech analysis, we propose the APGD feature as part of the whispering detection system.
- **MFCCs**, 40 coefficients including the zeroth, 128 mel bands, motivated by their shown universal applicability for audio analysis in various applications.
- **spectral centroid** and
- **spectral bandwidth**, shown to produce some improvement in the preliminary experiments.

With the all aforementioned features, the following frame-blocking parameters are used: window length 2048 samples, window hop 512 samples. The features are combined by means of frame-wise concatenation. The extraction of all the features except for APGD was performed using the LibROSA package [10].

Also the following features were studied within the framework: spectral rolloff, coefficients of fitting an polynomial to the columns of a spectrogram, zero crossing rate, chromagram, RMS energy. These have shown suboptimal results in the preliminary experiments and are excluded from the primary evaluation set up.

An option of incorporating context information into training is foreseen by means of stacking the features of the variable number of consecutive frames from the continuous signal of the same class. In addition to the added temporal information into training, this measure also aids in reducing the negative effect of pauses between words in the data, which are annotated as a part of larger-scale whispered segment.

C. Classifier

The primarily proposed classification methodology is based on deep neural networks. For each time frame t, a feature vector $\mathbf{x}_t$ is used as a learning instance for the DNN. The target output vector for each frame $\mathbf{y}_t$ is a binary vector with elements determined from the annotations as

$$y_t = \begin{cases} 1, & \text{whispering is observed in frame } t, \\ 0, & \text{no whispering is observed in frame } t. \end{cases} \quad (1)$$

With the help of the hierarchical topology of the DNN, high-level features are implicitly learned in the higher-level hidden layers. This makes it possible to model the vastly non-linear relationships between the input and the output.

The multi-label DNN architecture is composed of an input layer with the number of units equal to the input vector size, two or more hidden layers and an output layer with 2 units for the considered binary case. For each layer k, the outputs $\mathbf{h}^k$ are calculated from the weighted sum of the outputs from the previous layer $k-1$, starting from $\mathbf{h}^0 = \mathbf{x}$ and

$$\mathbf{g}^k = \mathbf{W}^k \mathbf{h}^{k-1} + \mathbf{b}^k, 1 \leq k < M, \quad (2)$$

$$\mathbf{h}^k = f(\mathbf{g}^k), \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{D \times S}$ is the weight matrix between layers $k-1$ and k, D and S are the number of units for layers $k-1$ and k respectively, $\mathbf{b} \in \mathbb{R}^S$ is the bias vector for layer k, $f(\cdot)$ is the nonlinear activation function for layer k and M is the total number of layers. For the hidden layer activation functions, `maxout` is used [11], and for the output layer activation function, softmax is used. Mean squared error cost function is chosen during the training of the DNN. In the classification stage, the prediction vector is binarised with an initial threshold 0.5 to obtain a binary detection estimation vector.

The implementation of the used DNN framework is a `Theano`-based [12] `pylearn2` library [13] with a `scikit-learn` [14]-compatible interface `scikit-neuralnetwork`.

With a prior assumption of the variations between whispered and normal speech happening relatively slowly, the possible short abrupt changes in the outputs of the DNN are safely eliminated by means of post-processing. A simple, median filtering -based approach is applied. For each binary detection estimation $z_t(l)$, the post-processed estimation $\tilde{z}_t(l)$ is obtained by taking a median of the previous 47 frames,

spanning 3 seconds. This effectively filters out the short bursts of detections.

The following parameters were set to the proposed systems. Two hidden layers (reasoned as being deep but not overly complicated) with 50 units each (selected as a reasonable value between the dimensions of the input and output) were learnt. Due to moderately large amount of training data, the learning rate was set to 0.001, and the training was terminated after 200 iterations. The validity of these values was briefly confirmed in preliminary tests on data, similar to the one used in evaluation, but different. The training of the system is computationally relatively light, possible to perform within 6 hours on a consumer-level desktop given at least 16 GB RAM.

Alternatively to the DNNs, such conventional classifiers as decision trees, Gaussian mixture model (GMM), random forest classifier, and Linear Support Vector Classification (LSVC) were included into the system as well, and their performance is compared to the DNNs' in the evaluation. After a brief preliminary parameter search, a value of 40 mixture components per class model has been set for training the GMMs. The random forest classifier consisted of ten trees.

III. Acoustic data

For this work, we collected a whispered vs. non-whispered speech corpus. It consists of four hours of audio recordings of diverse non-monotonous speech. The audio signals of 20 minutes each were obtained from 42 people pronouncing scripted English phrases in an dialogue scenario. The high quality audio signals were collected in a controlled noise-free low reverberation environment. We simulated telephone quality office speech by adding background noise artificially and encoding the signals with a speech codec.

As the text data, a set of modified Harvard sentences [15] was used. The sentences are phonetically balanced sample English phrases. The scripts were structured into six sets, each of which was divided into three identical subsets (for each vocal effort class: normal speech, loud whispering, quiet whispering), and each subset was further subdivided into two collections: 10 sentences where the first speaker asks a question and the other answers, and 10 sentences the other way around. The scripts contained markings of whether a sentence needs to be uttered normally, by whispering loudly (with the instruction: "as if you were whispering on the phone"), and whispered quietly ("as if you were whispering in quiet conditions"), accompanied by a note on the desired intonation (question, answer). The latter aspect is incorporated in order to facilitate naturalness of the emulated conversation.

The recording was performed as follows. Forty-two volunteers (18 females and 24 males) of various nationalities with confident English language skills were recruited. The recording was performed in a controlled environment: a noise-free acoustically treated room of dimensions $4.53 \times 3.96 \times 2.59$ m with low reverberation ($T_{60} = 0.26$ s). In each session, two participants with wearable microphones within 5 cm from the mouth were positioned within a distance of two meters from each other. Thereupon, they acted out the provided scripts, and the average target conversation was of length 20 minutes.

TABLE I
Measured average SDR values of the speech classes in the corpus with added office noise.

Speech type	Average SDR, dB
normal	-4.7
loud whispering	-5.0
quiet whispering	-7.9

Each speaker was recorded with two microphones: a lavalier condenser AKG C 520 and a clip-on Sennheiser MKE 2 P - C. The signals were fed into a RME UFX preamplifier and digitised in the format WAV PCM16, mono, 44.1kHz. The recordings were thereupon manually annotated in terms of timings of the uttered sentences and the corresponding vocal efforts.

Based on the clean corpus, its version with emulated real-life phone call conditions was generated by means of pre-processing. Randomly selected segments of background noise from the office environment from the IEEE AASP DCASE 2013 Challenge [16], Scene Classification task, were mixed into the recordings. The levels of both the noise and the clean data were kept original, so that the resulting signal-to-noise ratio is different for signals of different levels. This was done in order to preserve the differences of dynamics of whispered and non-whispered speech, which is in line with the expected real-life use case. The levels of added noise were perceptually confirmed to be adequate to the realistic case. Additionally, the phone line channel effects were emulated by means of GSM coder and decoder using the SoX implementation [17]. After adding the noise and channel distortions, the following average signal-to-distortion (SDR) ratios were measured for the whole corpus, see Table I.

The low levels of SDR are both due to the added noise and the significant degradation introduced by the GSM codec. The difference of SDR values between classes of speech was apparent also by perceptual investigation: the quietly whispered speech is indistinguishable under the noise floor. In the development stage, only the normal speech and loud whispering cases were therefore considered. A matched amount of purely background noise data is added to the dataset so that whispering is to be detected against both normal speech and pauses filled with background noise.

IV. Evaluation

This section presents the details of the evaluation of the proposed system with the metrics used, the details on the data splits as well as the obtained results.

A. Metrics and datasets

As evaluation metrics, two primary ones are considered: frame-wise and block-wise accuracy. For frame-wise evaluation, the performance of the system is measured as the number of correctly classified frames relative to the total number of

TABLE II
COMPARISON OF THE FEATURES EVALUATED WITH THE DECISION TREES CLASSIFIER.

Feature	Frame acc.	Block acc.	F1	Prec.	Rec.
APGD	**72.3**	91.7	62.7	0.65	0.60
MFCCs	71.0	88.2	60.3	0.62	0.59
APGD+MFCCs+centr.+bw	**73.5**	89.6	63.9	0.66	0.62
Spectral centroid	56.8	67.3	41.0	0.42	0.40
Spectral bandwidth	57.2	65.1	40.3	0.40	0.40
Spectral rolloff	65.4	68.0	40.9	0.34	0.52
Poly features	59.5	68.7	43.2	0.43	0.43
Zero crossing rate	64.1	64.4	8.0	0.04	0.46
Chromagram	57.3	70.3	43.7	0.46	0.41
RMSE	57.6	64.8	39.9	0.39	0.40

frames in the test subset. While being straightforward, this measure suffers from a natural drawback: when a sentence is annotated to be of continuously whispered speech, the pauses between the words are erroneously annotated, if considered frame-wise.

As an alternative, a block-wise measure is considered, which uses a sliding median window of length 3 seconds. Within a window, a majority vote over the predicted labels is performed, producing a single label for a block. The accuracy is then calculated as a ratio of correctly classified blocks to the total number of blocks within the test subset. Due to the use of median filtering, if the pauses are shorter than half of the block length, they are expectedly filtered out. Additionally, such measures as precision, recall and F1 score were used in the evaluation.

For a fair estimate of the performance of the system, the noisy and distorted version of the collected corpus is split into training and test subset while making sure that no recording of the same speaker is present in both subsets. That is, the division into training and test subsets is performed in terms of speakers. The gender proportion is preserved by means of performing splits with the same ratio for each gender separately. The split is done randomly, however, with a fixed seed in order to preserve comparability of the results given rapid development scenario. The proportions of the split are 70/30 for training/testing.

The evaluations were performed with the goals of obtaining optimal feature sets, their parameters, the comparison of classifiers and more thoroughly the parameters of the proposed DNN-based set up.

B. Features

The feature comparison evaluation was performed at the initial development stage, and decision trees were used as a classifier. The evaluated features were all the features used during the development of the system: all-pole group delay, MFCCs, spectral centroid, spectral bandwidth, spectral rolloff, coefficients of fitting an polynomial to the columns of a spectrogram, zero crossing rate, chromagram, RMS energy. The results are presented in Table II.

TABLE III
COMPARISON OF THE CLASSIFIERS WITH THE APGD+ MFCCS + CENTROID + BANDWIDTH FEATURES.

Feature	Frame acc.	Block acc.	F1	Prec.	Rec.
DNN	84.0	94.5	78.1	0.80	0.76
GMM	77.5	92.1	70.2	0.74	0.66
Decision tree	73.5	89.6	63.9	0.66	0.62
Random forest	79.9	87.0	68.8	0.62	0.77
SVM, linear kernel	80.6	88.0	69.7	0.63	0.79

Based on these experiments, we select the APGD + MFCCs + centroid + bandwidth feature combination, which manages to incorporate the important spectral information both from magnitude and phase points of view. Further evaluations are performed with this fixed feature set.

C. Classifiers

The comparison was performed between the conventional (decision trees, GMM, SVM) and state-of-the-art (DNN) classifiers. As features, the best performing combination in terms of frame-wise accuracy was selected based on the feature-wise evaluations. For DNN, the initial architecture of two hidden layers with 50 units each and no context-aware training was used.

The evaluation results are presented in Table III. Indeed, even though the dataset is only moderately large, the proposed DNN architecture manages to outperform the conventional classifiers. Further experiments were performed within the DNN set up in order to find number of hidden units, and values of 30-50 neurons per hidden layer within the two hidden layers set up were shown to perform best.

The value also is in line with the popular rule-of-thumb for selecting number of hidden units between the dimension of the features (72) and dimension of the output. For the final implementation, the value for the number of neurons in each of the hidden layers is set to 50.

D. Context-aware training of the DNN

The effect of incorporating context information by means of concatenating consecutive frames was studied in this experiments. The size of context is set in frames, each of hop 64 ms given the sampling rate (8 kHz) and frame blocking (hop 512 samples) of the noisy and corrupted version of the dataset. The evaluation results with different context lengths are presented in Table IV.

We observe a noticeable performance boost when increasing the span of the context in the training. A possible explanation is the fact, that the training data contains pauses between words, which are coarsely annotated as whispering. When incorporating several frames into a training vector, the chances of whispering being present somewhere in that sample increase, therefore the classifier is more likely to get trained on data which always contains some whispering. For the non-whispering class, this does not hold, since the corresponding data certainly does not contain any whispering whatsoever.

TABLE IV
EFFECT OF INCORPORATING CONTEXT BY MEANS OF CONSECUTIVE FRAMES INTO THE TRAINING OF DNN.

Context size, frames	Frame acc.	Block acc.	F1	Prec.	Rec.
0	84.0	94.5	78.1	0.80	0.76
5	87.4	94.1	82.2	0.81	0.83
10	89.4	93.5	85.6	0.88	0.83
20	91.7	94.6	88.5	0.89	0.88
30	91.8	93.8	88.6	0.90	0.88

From these observations we conclude, that a context span of a length comparable to the expected length of a whispered phrase is a reasonable choice. Thirty frames with the default set up stand for approximately two seconds of speech and is expectedly a good value. Given sufficient resources, setting the context span up to 30 frames is reasonable, however, further increase stands for less realistic long sentences of whispering.

V. CONCLUSIONS

A DNN-based whispered speech detection system was implemented. A set of optimal phase- and magnitude-based features was proposed. A moderately large corpus of whispered and non-whispered speech was collected. A network architecture, capable of successfully incorporating large amounts of training data and achieving good generalisation capabilities and real-life functionality has been proposed. The extensive evaluations of the intermediate steps of the system as well as the final set up justify the proposed architecture and show impressive performance results of frame-wise accuracy 91.8%.

Further improvements in the future are foreseen. Most importantly, incorporating the current state-of-the-art of neural computation, namely, LSTM, as well as perform data augmentation to further increase the amount and diversity of training data appears worthwhile.

REFERENCES

[1] K. Yatani and K. N. Truong, "Bodyscope: A wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 341–350. [Online]. Available: http://doi.acm.org/10.1145/2370216.2370269

[2] P. Zelinka and M. Sigmund, "Automatic vocal effort detection for reliable speech recognition," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, Aug 2010, pp. 349–354.

[3] M. A. Carlin, B. Y. Smolenski, and S. J. Wenndt, "Unsupervised detection of whispered speech in the presence of normal phonation." in *INTERSPEECH*, 2006.

[4] P. Rajan, T. Kinnunen, C. Hanili, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proc. Interspeech 2013*, 2013, pp. 2489–2493.

[5] A. Diment, E. Cakir, T. Heittola, and T. Virtanen, "Automatic recognition of environmental sound events using all-pole group delay features," in *European Signal Processing Conference 2015 (EUSIPCO 2015)*, Nice, France, Aug. 2015.

[6] J. Pohjalainen, T. Raitio, H. Pulakka, and P. Alku, "Automatic detection of high vocal effort in telephone speech." in *INTERSPEECH*, 2012, pp. 691–694.

[7] M. F. Schwartz, "Power spectral density measurements of oral and whispered speech," *Journal of Speech, Language, and Hearing Research*, vol. 13, no. 2, pp. 445–446, 1970.

[8] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[9] J. Sundberg, R. Scherer, M. Hess, and F. M "Whispering — single-subject study of glottal configuration and aerodynamics," *Journal of Voice*, vol. 24, no. 5, pp. 574 – 584, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089219970900006X

[10] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015.

[11] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 1319–1327.

[12] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[13] I. J. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien, and Y. Bengio, "Pylearn2: a machine learning research library," *arXiv preprint arXiv:1308.4214*, 2013. [Online]. Available: http://arxiv.org/abs/1308.4214

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "Ieee recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[16] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "IEEE AASP challenge: Detection and classification of acoustic scenes and events," Technical Report, Queen Mary University of London, Tech. Rep., 2013.

[17] (2015) SoX - Sound eXchange. Software, Rev. 14.4.2. [Online]. Available: http://sox.sourceforge.net/