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

Dominant Distortion Classification for Pre-Processing of Vowels in Remote Biomedical Voice Analysis

Amir Hossein Poorjam [1] *, Jesper Rindom Jensen* [1]*, Max A. Little* [2] *and Mads Græsbøll Christensen* [1]

[1] Audio Analysis Lab, AD:MT, Aalborg University, Aalborg, DK
[2] Engineering and Applied Science, Aston University, Birmingham, UK
[2] Media Lab, MIT, Cambridge, Massachusetts, USA
[1] {ahp,jrj,mgc}@create.aau.dk , [2] max.little@aston.ac.uk

Abstract

Advances in speech signal analysis facilitate the development of techniques for remote biomedical voice assessment. However, the performance of these techniques is affected by noise and distortion in signals. In this paper, we focus on the vowel /a/ as the most widely-used voice signal for pathological voice assessments and investigate the impact of four major types of distortion that are commonly present during recording or transmission in voice analysis, namely: background noise, reverberation, clipping and compression, on Mel-frequency cepstral coefficients (MFCCs) – the most widely-used features in biomedical voice analysis. Then, we propose a new distortion classification approach to detect the most dominant distortion in such voice signals. The proposed method involves MFCCs as frame-level features and a support vector machine as classifier to detect the presence and type of distortion in frames of a given voice signal. Experimental results obtained from the healthy and Parkinson's voices show the effectiveness of the proposed approach in distortion detection and classification.

Index Terms: distortion analysis, MFCC, remote biomedical voice assessment, support vector machine

1. Introduction

Sustained vowels are widely used for evaluation of pathological voice caused by a range of medical disorders. Vowels have two main advantages: first, the complexity of modeling articulatory movement during running speech is avoided [1], and second, experimental studies show that most dysphonic speakers cannot produce steady, sustained vowel sounds [2]. Among vowels, the vowel /a/ is sufficient for many voice analysis applications [3], [4]. During production of the vowel /a/, the vocal tract is more open than other vowels resulting in minimal air pulse reflections between the vocal tract and the vocal folds [1]. Using clean and sustained /a/ vowels, Tsanas et al. [4] achieved almost 99% overall accuracy in detecting Parkinson's disease (PD) from voice recordings, for example.

Due to advances in automatic voice analysis, remote voice assessment is becoming feasible [5], [6]. For example, recently smartphones are being investigated as tools for measuring pathological voice [7] since smartphones are ubiquitous and inexpensive devices with built-in, high-quality microphones. Compared to samples recorded in a clinic or a sound booth, recordings from smartphones in most environments are subject to many types of linear and nonlinear distortion. The presence of distortion in voice signals degrades the performance of algorithms designed to quantify medical symptoms from voice recordings [8]. In particular, the performance of different algorithms for PD detection under a variety of acoustic conditions has been evaluated in [9] and it has been demonstrated that background noise and the use of codecs significantly degrade detection performance.

Several approaches to detect different types of noise and distortion in voice signals have been studied, most of have focused on detecting a single and specific kind of distortion in voice [10–14]. In this study, we consider the vowel /a/ and aim to detect four different types of noise and distortion that are commonly present during recording or transmission in remote voice analysis, namely: background noise, room reverberation, peak clipping and coding (i.e. speech compression). Although there are an infinite number of possible levels, types and combinations of distortion in real-world scenarios, this study aims to provide a simplified approach to detect the most dominant distortion in the signal, which would be useful in practical applications where it is important to know whether a frame is distortion free or needs enhancement. We assume that if a given frame is considered as corrupted, there is a specific type of noise or distortion which dominates over other distortions. Following this, we investigate the behavior of Mel-frequency cepstral coefficients (MFCCs, widely-used features in voice-based biomedical applications [8], [15]) in the presence of the four kinds of distortion and noise. Then, a new method is proposed which uses a support vector machine (SVM) as classifier and MFCCs as features for that classifier, to detect distortion in each frame. MFCCs are selected because of their sensitivity to changes in signal characteristics due to noise, distortions or articulatory movements [16].

2. Effects of distortion on MFCCs

The proposed method is based upon experimental observations of the effect of distortions on MFCCs reported next. This experimental analysis reveals that different levels and types of distortion cause MFCCs to shift to different regions of the space spanned by the MFCC values, and changes the covariance of these values. To explore this effect, we take successive frames from the center of the clean vowel /a/ uttered by 45 healthy speakers and extract the MFCCs under different types and levels of distortion and noise. We then evaluate the shift in the sample mean and covariances of the MFCCs computed on the distorted signals.

2.1. MFCC Features

MFCCs are based on the source-filter theory of speech production [17]. To compute MFCCs, we take the discrete Fourier transform (DFT) of the speech frames. Then, the power

This work was funded by the Danish Council for Independent Research, grant ID: DFF 4184-00056

http://dx.doi.org/10.21437/Interspeech.2017-378

spectrum is computed and passed through a set of triangular filter banks, linearly spaced on the Mel-frequency scale. The log-energy output of the filter bank, which is sensitive to small changes in signal characteristics due to noise, distortions or articulatory movements [16], is then passed through the discrete cosine transform (DCT). The MFCCs are the amplitudes of the DCT coefficients. Specifically, the p^{th} MFCC coefficient of the k^{th} frame is calculated as [8]:

$$\phi_k[p] = \frac{1}{M+1}\sum_{q=1}^{M}\log|\tilde{S}_k(q)|\cos\left(\frac{\pi q}{M+1}p\right), \quad (1)$$

where M is the number of Mel-band filters and $\tilde{S}_k(q)$ is the estimate of the spectral energy in the q^{th} band calculated as:

$$\tilde{S}_k(q) = \sum_{i\ni f_i^{\text{Mel}}\in I_q^{\text{Mel}}}\left(1-\frac{|f_i^{\text{Mel}}-\frac{q}{M+1}F^{\text{Mel}}|}{\frac{\Delta f^{\text{Mel}}}{2}}\right)|S_k(i)|, \quad (2)$$

where $I_q^{\text{Mel}} = [\frac{q-1}{M+1}F^{\text{Mel}}, \frac{q+1}{M+1}F^{\text{Mel}}]$ is the q^{th} filter band in Mel-frequency scale, f_i^{Mel} is the i^{th} Mel frequency, F^{Mel} is the maximum frequency in the Mel domain, $\Delta f^{\text{Mel}}/2$ is the width of the Mel bands and S_k is the short-time DFT of the k^{th} frame. The transformation from the linear domain to the Mel domain is performed by [18]:

$$f^{\text{Mel}} = \frac{1000}{\log_{10}2}\log_{10}\left(1+\frac{f}{1000}\right). \quad (3)$$

In this study, 13 MFCC coefficients are extracted for each frame. In addition, *delta* and *double-delta* coefficients, defined as the first- and second-order time-differences of the MFCC coefficients which capture the dynamic changes between frames, are appended to the MFCCs to form a 39-dimensional vector.

Considering (1) – (3), the effects of distortions on MFCCs are complex since during the MFCC calculations, a corrupted signal passes through several nonlinear functions. These effects can even be more complex when a signal has been subject to a nonlinear distortion such as clipping or compression. To evaluate the behavior of MFCCs in the presence of noise and distortion, we take successive 30 ms long frames of the vowel /a/ uttered by 45 healthy speakers and compute the change in the covariance matrix and the mean of the MFCCs under different types and levels of distortion. Specifically, the mean shift can be defined as:

$$\xi(j) = \frac{1}{N}\sum_{n=1}^{N}\parallel\boldsymbol{\mu}_n^{\text{d}j}-\boldsymbol{\mu}_n^{\text{c}}\parallel_2, \quad (4)$$

where N is the number of speakers, $\|\cdot\|_2$ represents the 2-norm, and $\boldsymbol{\mu}_n^{\text{c}}$ and $\boldsymbol{\mu}_n^{\text{d}j}$ are the means of the MFCCs computed respectively from clean signal and distorted signals from the n^{th} speaker subject to the j^{th} distortion level. $\xi = 0$ indicates that the mean of the corrupted MFCCs is unchanged in feature space. The larger the value of ξ, the farther the MFCC vector is moved with respect to the clean one. The change in the covariance matrix of the MFCC under the j^{th} distortion level is measured as:

$$\delta(j) = \frac{1}{N}\sum_{n=1}^{N}\frac{\text{tr}\left(\boldsymbol{\Sigma}_n^{\text{d}j}\right)}{\text{tr}\left(\boldsymbol{\Sigma}_n^{\text{c}}\right)}, \quad (5)$$

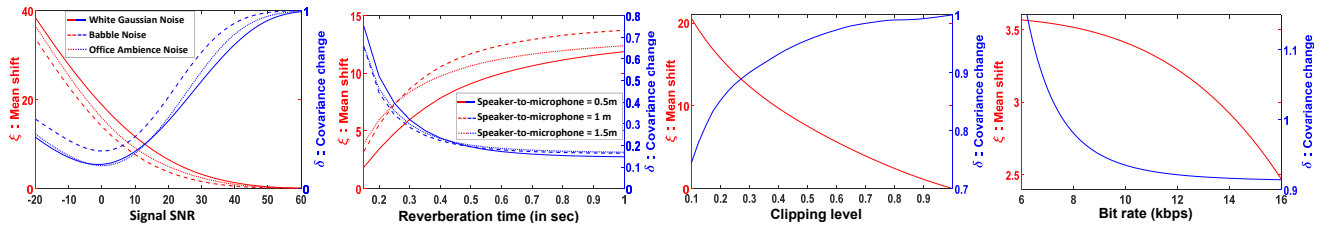
where $\boldsymbol{\Sigma}_n^{\text{c}}$ and $\boldsymbol{\Sigma}_n^{\text{d}j}$ are respectively the covariance matrices of the MFCCs extracted from the clean and corrupted utterances of the n^{th} speaker, and $\text{tr}(\cdot)$ is the *trace* operator that maps the MFCC covariance matrix to a single real number which represents the sum of variances for individual dimensions of the MFCC vector. $\delta = 1$ represents no change in covariance. A value of $\delta < 1$ indicates a reduction in covariance with respect to the covariance of the clean MFCC. That is, the MFCCs become more compact in the feature space.

2.2. Impact of different distortions on MFCCs

In the first experiment, we investigate the impact of background noise on MFCCs by corrupting clean vowels /a/ uttered by 45 healthy speakers by three commonly-encountered environmental noise types, namely "white Gaussian noise", "quiet office ambience noise" and "babble noise" under different signal-to-noise ratio (SNR) conditions (ranging from -20 dB to 60 dB in 1 dB steps). Babble noise, which consists of multiple speakers talking in the background, has rapid, time-evolving structure and is considered a challenging type of noise in many speech-based applications due to its similarity to the target speech [19]. The office environment noise represents a general atmosphere of a medium size room including the sound of air conditioning systems and very weak background noise from outside. Figure 1(a) shows the impact of different types and levels of noise on the mean and the covariance matrix of MFCCs. The left vertical axis represents the amount of mean shift as defined in (4) and the right vertical axis represents the relative change in the covariance matrix as defined in (5). The plot suggests that variable noise levels shift the mean of MFCCs to different, but predictable, regions in the feature space. It can be observed that the amount of shift monotonically increases as the level of noise increases. Moreover, it can be noticed that the covariance of the noisy MFCCs is always smaller than that of the clean one. However, the covariance does not monotonically reduce. This is probably due to the fact that as the SNR goes below 0 dB, the noise dominates the signal and the MFCCs take on a different profile.

Reverberation in voice recordings is caused by superimposed reflections of the original sound wave coming from different surfaces in an acoustic environment and is known to have a detrimental impact on numerous signal processing tasks. To study the effect of reverberation on MFCCs, we filtered the clean signal with synthetic room impulse responses (RIRs) of reverberation times (RTs) varying from 150 ms to 1 s measured at a room of dimension 5m×4m×3m. Furthermore, to evaluate the effect of different source-to-receiver distances on the MFCCs, the experiments are repeated with three different speaker-to-microphone distances, namely 0.5 m, 1 m and 1.5 m. The RIRs are generated using the image method [20] which is implemented using the RIR Generator toolbox [21] in MATLAB. Figure 1(b) illustrates similar trends for MFCCs under different speaker-to-microphone distances. The mean shift increases as the RT increases. We can observe that when the microphone records from a close distance, the amount of shift is always smaller than when the microphone is recording from a larger distances from the speaker. For large speaker-to-microphone distances, however, we observe a different trend as the RT exceeds 250 ms. Reverberation reduces the covariance of the MFCCs as the RT increases.

Peak clipping and speech coding are two common nonlinear speech signal modifications. Peak clipping occurs when the amplitude of a speech signal exceeds the dynamic range of the analogue-to-digital converter which introduces nonlinear distortion into the signal and affects the subjective quality of



(a) *Background noise* (b) *Reverberation* (c) *Peak clipping* (d) *Compression*

Figure 1: *Impact of different types and levels of distortion on the mean and covariance matrix of the MFCCs. The left vertical axes represent ξ defined in (4) which is the amount of mean shift. $\xi = 0$ indicates that the mean of the corrupted MFCCs is not shifted in the feature space. The larger the value of ξ, the farther the MFCC vector is positioned with respect to the clean one. The right vertical axes represent δ defined in (5) which is the relative change in the covariance matrix. $\delta = 1$ indicates no change in covariance of the corrupted MFCCs, $\delta > 1$ represents increase in covariance with respect to the covariance of the clean MFCCs and $\delta < 1$ indicates that the MFCCs become more compact in the feature space.*

speech [22]. On the other hand, communication channels typically use lossy codecs such as code-excited linear prediction (CELP) to compress speech signals to lower bit rates, which inevitably degrades the quality of the speech [23]. To study the effect of peak clipping on MFCCs, we define the clipping level as a proportion of the unclipped peak absolute signal amplitude to which samples greater than this threshold are limited. The clean recordings of the vowel /a/ are clipped with clipping levels varying from 0.1 to 1 in 0.025 steps. Figure 1 (c) illustrates the impact of peak clipping on the MFCCs. As the clipping level increases, the mean of the MFCCs is positioned farther away from that of the clean signal. MFCCs of a clipped signal possess smaller covariance matrix values compared to that of the clean MFCCs and become smaller as the clipping level increases. To investigate the behavior of MFCCs when a speech signal has undergone the distortion of a speech codec, the clean vowels /a/ are coded by a CELP codec with three different standard bit rates, namely 6.3, 9.6 and 16 kbps [24]. Figure 1 (d) shows the impact of speech compression on MFCCs. The plot is produced by fitting a second order power function to the calculated ξ and δ as:

$$\tilde{\xi}(j) = -2.35 \times 10^{-5} \times j^{3.88} + 3.59 \quad (6)$$

$$\tilde{\delta}(j) = 1797 \times j^{-4.89} + 0.91 \quad (7)$$

We can observe that speech compression shifts the MFCCs to a farther position (with respect to the position of the clean ones) as the compression rate increases. On the other hand, although MFCCs of a voice signal coded at 16 kbps and 9.6 kbps possess smaller covariance matrices compared to the covariance of the clean one, we observe a larger covariance than that of the clean MFCCs when a signal is coded at 6.3 kbps. The empirical curve fitted to δ also suggests that MFCCs of a signal compressed at 7.3 kbps are expected to have a comparable covariance matrix with respect to the covariance of the clean MFCCs.

3. The proposed distortion classification system

Motivated by the experimental findings above, we introduce a new method for noise and distortion classification to detect the presence and type of noise/distortion in /a/ vowels. Although a recording can be subject to an infinite number of possible types and levels of noise and distortion in real scenarios, our approach focuses on detecting the most dominant corruption present in any frame. We assume the simplifying model that if a given

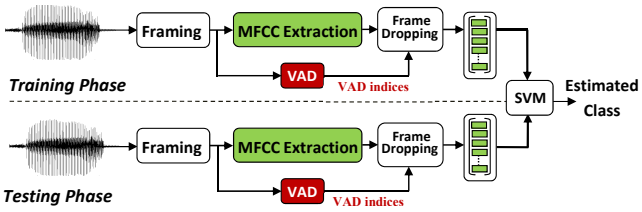


Figure 2: *Block diagram of the proposed method for distortion/noise classification, training and testing phases.*

frame of a voice recording is corrupted, there is a single type of noise or distortion which dominates. The block diagram of the proposed approach in training and testing phases is illustrated in Figure 2. Using a Hamming window, recordings are segmented into frames of 30 ms. For each frame of a vowel (which can be clean or corrupted), a 39-dimensional MFCC vector is computed. Using an energy-based voice activity detection algorithm [25], silent frames at the beginning and the end of the signals are excluded. Then, a multiclass SVM with a radial basis function kernel estimated on the training frames is used to classify distortion in an unseen frame during testing. Introduced by Vapnik et.al [26], SVMs are powerful discriminative pattern classifiers which find an optimal separating hyperplane in a high dimensional nonlinear feature space formed using kernels applied to the input feature space.

4. Experimental setup

The proposed system for distortion/noise recognition in /a/ vowels was developed and validated using two different databases. The first database consists speech samples of healthy speakers. This database contains different clean vowels uttered by 45 men, 48 women and 46 childeren, recorded by a dynamic microphone, sampled at 16 kHz and range from 370 ms to 780 ms long [27]. There is no dysphonia variability. The only uncontrolled parameter is the speaker variability. From this database, we have chosen 93 samples of /a/ vowels produced by 45 male and 48 female speakers. Furthermore, to evaluate the proposed system with more realistic pathological voice signals, we used a PD voice database since the vast majority of people with PD exhibit some form of vocal disorder [28]. This database was generated through collaboration between Sage Bionetworks, PatientsLikeMe and Dr. Max Little as part of the Patient Voice

Table 1: *Frame- and recording-level classification performance for the healthy voice and the Parkinson's voice databases in the form mean ± STD computed using a 5-fold CV.*

Database	**Frame-Level Classification Accuracy (in % ± STD)**						**Recording-Level Classification Accuracy (in % ± STD)**					
	Clean	Noisy	Clipped	Coded	Reverb.	Overall	Clean	Noisy	Clipped	Coded	Reverb.	Overall
Healthy voice	61±6	92±3	82±3	71±6	85±4	78±1	77±12	100±0	98±3	82±11	90±7	89±4
Parkinson's voice	48±5	89±3	74±6	77±8	66±5	72±4	55±11	97±4	82±7	85±9	77±4	79±3

Analysis study (PVA)[1]. The samples of this database are the telephone recordings of the sustained vowels /a/ produced by 779 PD patients of both genders, sampled at 8 kHz and range from 3 s to 30 s long. From this database, we randomly selected 48 female and 26 male samples of 7 s to 15 s duration. Then, we used 3 s of the middle of the signals, where the speakers produced a steady sustained vowel. This database has both speaker- and dysphonia- variability. Moreover, the recordings may have already some types of distortion such as background noise and reverberation or may have been through one or more codec since they are collected over the telephone network, which makes the noise/distortion classification more challenging.

To create a database for distortion/noise detection, we enlarged the databases by adding the distorted versions of all recordings by applying different types and levels of noise and distortion which typically present in the recordings of remote voice analysis. Specifically, for noise, we added "babble", "white Gaussian" and "office ambiance" noises at 15 dB, 10 dB and 5 dB. For peak clipping, the clipping level was set to 0.3, 0.4, 0.5 and 0.6. Signals were compressed using 6.3 kbps, 9.6 kbps and 16 kbps CELP codecs. To provide reverberant signals, recordings were filtered by 8 different real RIRs of the AIR database [29]. The RIRs are measured with mock-up phone in hand-held and hands-free positions in four realistic indoor environments, namely an office, a lecture room, a corridor and a stairway. The measured RTs range from 390 ms to 1.47 s [30]. Then, using a Hamming window of length 30 ms, we created a database of 30 ms clean and corrupted frames for both databases. The recordings of each database are then divided into two subsets: a training subset consisting of 80% of the speakers, and a testing subset consisting of 20% of the speakers. The resulting training and testing subsets of the enlarged healthy vowel database consist of 5105 and 1360 frames, respectively. The training and the testing subsets of the enlarged PD voice database consist of 30150 and 7800 frames, respectively. The enlarged databases have the same number of frames per class of noise/distortion.

To detect different types of distortion in a given frame, a multiclass SVM classifier implemented in the LIBSVM toolbox [31] in MATLAB is used. The hyper-parameters of the SVM, namely the RBF kernel spread and SVM regularization parameter, were selected by 5-fold cross-validation (CV) on 10% of the training data assigned as the tuning subset.

5. Results and discussion

We used 5-fold CV to evaluate the classification performance in terms of the number of correctly classified test frames. The results over all CV repetitions using healthy and the PD voices are reported in the first and the second rows of Table 1, respectively. Assuming that the most dominant distortion in an utterance usually affects the majority of frames, we also extend the proposed method to the recording-level by applying a majority voting algorithm over all frames of a signal. The table reports the classification accuracy both at frame and recordings levels in the form mean ± one standard deviation. The reported numbers for different classes are the diagonal elements of the corresponding confusion matrices and the last columns report the overall classification accuracy.

The effectiveness of MFCCs in distortion classification (particularly for noisy frames) can be observed. The results for healthy voices are consistent with the behavior of MFCCs in the presence of different types and levels of distortion observed in Section 2.2. Considering Figure 1, MFCCs of the coded signals are, on average, positioned closer to the MFCCs of the clean signals, while noise, clipping and reverberation shift the MFCCs farther away from the position of clean MFCCs. Moreover, the covariance of MFCCs extracted from coded signals is comparable to that of the clean signal, while the MFCCs for noisy, reverberant and clipped signals are more compact in the feature space. Taking these two observations into account, the MFCCs of the coded and clean frames are more likely to be overlapping in the feature space which results in misclassification, particularly when there is speaker variability in the data.

Although the proposed method is effective in distortion classification for both healthy and pathological voices, we observe a degradation in overall classification performance (particularly for clean frame detection) when the system is evaluated using the PD voice database. The first factor affecting the results is the dysphonia variability in the PD voice database since the presence of pathologies in speech is related to signal variability. Moreover, bearing in mind that the recordings in the PD voice database have been collected over the telephone, these signals may have already been through one or more codecs. This means that some coded frames have been presented to the classifier as "clean" ones during the training phase which will result in some classification performance degradation.

6. Conclusions

In this study, the impact of four major types of distortion, namely background noise, reverberation, clipping and speech compression on MFCCs of the frames of the vowel /a/ has been analyzed. These distortions are commonly present in voice signals during recording or transmission in remote pathological voice assessments. It has been demonstrated experimentally that introducing different types and levels of distortion to the vowel results in predictable changes in mean and covariance matrix of the MFCCs. Motivated by this observation, a new approach for detecting the dominant type of distortion is proposed, which uses MFCCs as frame-level acoustic features and an SVM as the classifier. Experimental results using recordings of healthy speakers and speakers with PD (as an example of people with voice disorders) show the effectiveness of the proposed system in distortion classification. Since the presence of disorders in speech is closely related to signal variability, a slight degradation in classification performance has been observed when the PD voices were analyzed.

[1]They were obtained through Synapse ID [syn2321745]

7. References

[1] I. Titze, *Principles of voice production*, 2nd ed. Iowa City: National Center for Voice and Speech, 1999.

[2] J. Schoentgen and R. De Guchteneere, "Time series analysis of jitter," *J. Phon.*, vol. 73, pp. 189–201, 1995.

[3] F. Klingholtz, "Acoustic recognition of voice disorders: a comparative study of running speech versus sustained vowels." *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2218–24, may 1990.

[4] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 1264–1271, 2012.

[5] R. J. Moran, R. B. Reilly, P. De Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 468–477, 2006.

[6] P. A. Mashima and C. R. Doarn, "Overview of telehealth activities in speech-language pathology," *Telemed. e-Health*, vol. 14, no. 10, pp. 1101–1117, dec 2008.

[7] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Relat. Disord.*, vol. 21, no. 6, pp. 650–653, 2015.

[8] R. Fraile, N.Sáenz-Lechón, J. I. Godino-Llorente, and V. Osma-Ruiz, "Use of Mel frequency cepstral coefficients for automatic pathology detection on sustained vowel phonations: mathematical and statistical justification," in *4th Int. Symp. Image/Video Commun. over Fixed Mob. Networks*, no. 3, 2008.

[9] J. Vasquez-Correa, J. Serra, J. F. Orozco-Arroyave, J.R. Vargas-Bonilla, and E. Noth, "Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech," in *ICASSP*, 2017, pp. 5065–5069.

[10] W. Yuan and B. Xia, "A speech enhancement approach based on noise classification," *Appl. Acoust.*, vol. 96, pp. 11–19, 2015.

[11] K. El-maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *ICASSP*, pp. 237–240, 1999.

[12] S. Aleinik and Y. Matveev, "Detection of clipped fragments in speech signals," *Int. J. Electr. Comput. Energ. Electron. Commun. Eng.*, vol. 8, no. 2, pp. 286–292, 2014.

[13] J. Eaton and P. A. Naylor, "Noise-robust detection of peak-clipping in decoded speech," *ICASSP*, pp. 7019–7023, 2014.

[14] J. M. Desmond, L. M. Collins, and C. S. Throckmorton, "Using channel-specific statistical models to detect reverberation in cochlear implant stimuli." *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1112–20, 2013.

[15] A. Dibazar, S. Narayanan, and T. Berger, "Feature analysis for automatic detection of pathological speech," in *Second Jt. EMBS/BMES Conf.*, vol. 1, 2002, pp. 182–183.

[16] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, 2012.

[17] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 2nd ed. New York: IEEE Press, 2000.

[18] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics*. Kluwer Academic Publishers, 1999.

[19] N. Krishnamurthy and J. Hansen, "Babble noise: modeling, analysis, and applications," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 17, no. 7, pp. 1394–1407, 2009.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, p. 943, 1979.

[21] E. A. P. Habets, "Room impulse response generator," International Audio Laboratories Erlangen, Tech. Rep., 2010.

[22] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. Commun.*, vol. 33, no. 8, pp. 801–808, 1985.

[23] P. E. Souza, "Effects of compression on speech acoustics, intelligibility, and sound quality." *Trends Amplif.*, vol. 6, no. 4, pp. 131–65, 2002.

[24] R. Goldberg and L. Riek, *A practical handbook of speech coders*. CRC Press, 2000.

[25] T. Kinnunen and H. Li, "An Overview of Text-Independent Speak er Recognition : from Features to Supervectors," *Speech Commun.*, vol. 1, 2009.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[27] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, 1995. [Online]. Available: http://homepages.wmich.edu/∼hillenbr/voweldata.html

[28] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease." *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998.

[29] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.

[30] M. Jeub, M. Schäfer, and H. Krüger, "Do we need dereverberation for hand-held telephony?" in *Int. Congr. Acoust.*, 2010, pp. 1–7.

[31] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," National Taiwan University, Tech. Rep., 2016.