SELECTING DISORDER-SPECIFIC FEATURES FOR SPEECH PATHOLOGY FINGERPRINTING

Visar Berisha[1]*, Steven Sandoval*[2]*, Rene Utianski*[1]*, Julie Liss*[1]*, and Andreas Spanias*[2]

Arizona State University
[1]Department of Speech and Hearing Science, [2]School of ECEE, SenSIP Center
Tempe, AZ 85287

ABSTRACT

The general aim of this work is to learn a unique statistical signature for the state of a particular speech pathology. We pose this as a speaker identification problem for dysarthric individuals. To that end, we propose a novel algorithm for feature selection that aims to minimize the effects of speaker-specific features (e.g., fundamental frequency) and maximize the effects of pathology-specific features (e.g., vocal tract distortions and speech rhythm). We derive a cost function for optimizing feature selection that simultaneously trades off between these two competing criteria. Furthermore, we develop an efficient algorithm that optimizes this cost function and test the algorithm on a set of 34 dysarthric and 13 healthy speakers. Results show that the proposed method yields a set of features related to the speech disorder and not an individual's speaking style. When compared to other feature-selection algorithms, the proposed approach results in an improvement in a disorder fingerprinting task by selecting features that are specific to the disorder.

Index Terms— speech pathology, dysarthria, machine learning, feature selection

1. INTRODUCTION

Intelligibility of patients with speech pathologies is currently assessed through subjective tests performed by trained speech-language pathologists. Subjective tests, however, tend to be inconsistent, costly and, oftentimes, not repeatable. In fact, research has shown poor inter- and intra-rater reliability in clinical assessment. Furthermore, clinicians working with patients form a bias based on their interactions, resulting in intelligibility assessment of limited validity and reliability [1–4]. The goal of our work is to augment speech language pathologists with a digital signature of an individual's speech pathology state. This signature can then be tracked over time to assess the efficacy of the provided treatment, or the progression of a disease state. In this paper, we propose a novel feature selection algorithm that identifies a series of pathology-specific features while attempting to minimize speaker-specific effects.

We pose this problem as a speaker identification problem for dysarthric individuals. Principal differences between different speakers arise from differences in their speaking style and differences in the speech manifestation of their neurological disorder. We propose an algorithm that selects features that mostly contribute to the differences in speech pathology rather than speaking style. More specifically, we select acoustic features that discriminate well between dysarthric individuals, but provide minimal separation between healthy speakers. This results in a feature set that focuses on acoustic cues capturing differences in disorder rather than differences in speaking style. Consider the following example illustrated by Fig. 1 and Fig. 2 as motivation, which shows the distribution of two features (pitch period and pairwise energy variability) for 2 dysarthric and 2 healthy speakers, respectively. It is clear from Fig. 1 that the pitch period provides significant separation between the two dysarthric speakers, however it also provides separation between the healthy speakers. Alternatively, as is apparent from Fig. 2, the pairwise frame energy variability provides significant separation between the two dysarthric speakers, but has significant overlap for the healthy speakers. This is a feature that analyzes the variability in energy between consecutive 20 ms frames in the octave band centered at 2 kHz. With this, we conclude the pairwise energy variability measures are sensitive to certain aspects of the speech pathology, whereas the pitch metrics are sensitive to speaking style. The general aim of this work is to extend this example by designing an algorithm that selects features that are more sensitive to certain aspects of speech pathology rather than speaking style.

The literature contains limited work in this area. In [5], the au-

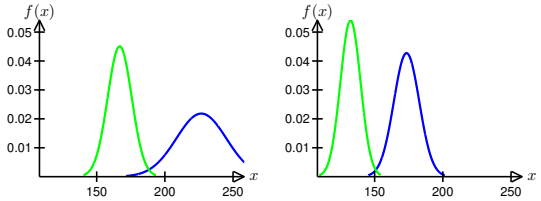

Fig. 1. The PDF of the *average pitch (Hz)* modeled as a normal distribution for (a) two dysarthric speakers (b) two healthy speakers.

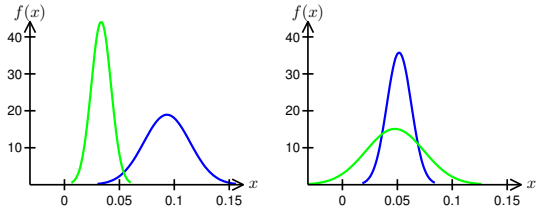

Fig. 2. The PDF of the *Pairwise Variability at 2000 Hz* modeled as a normal distribution for (a) two dysarthric speakers (b) two healthy speakers.

This research was supported in part by National Institute of Health, National Institute on Deafness and Other Communicative Disorders grants 2R01DC006859 (J. Liss) and 1R21DC012558 (J. Liss and V. Berisha).

thors rely on rhythm metrics estimated through envelope modulation spectra to classify between different dysarthria types. In [6] the authors make use of acoustic cues to detect Parkinson's disease using only speech. In [7–9] the authors develop an algorithm for assessing intelligibility using a regression scheme that makes use of a number of acoustic cues. In [9–12] the authors present a number of schemes for assessing speech quality and intelligibility by comparing to a clean reference signal. This paper is fundamentally different from the previous work, as we are not interested in discriminating between dysarthria types or predicting intelligibility. The goal here is to identify a set of features that act as a signature for the state of a speech pathology in an individual. Toward this end, we derive a cost function for selection of features and develop an efficient algorithm for solving it (Section 2). Following, we provide comparative results and show the efficacy of our proposed technique (Section 3). In Section 4, we provide concluding remarks.

2. FEATURE SELECTION

The goal of our research here is to design an algorithm which takes a large set of features and selects a subset of features that act as a signature for the state of a speech pathology. In order to develop this algorithm, we next introduce our feature sets, define our cost function, and define an efficient algorithm for solving it.

2.1. Feature Description

Although Mel-Frequency Cepstral Coefficients (MFCCs) have been the prevalent feature used for automatic speech recognition for over 30 years [13, 14], our feature set is composed of three different types of features more commonly found in the study of pathological speech: Envelope Modulation Spectra (EMS) features, Long-Term Average Spectra (LTAS) features, and ITU P.563 features. We discuss these below.

EMS - The envelope modulation spectrum (EMS) is a representation of the slow amplitude modulations in a signal and the distribution of energy in the amplitude fluctuations across designated frequencies, collapsed over time [5]. It has been shown to be a useful indicator of atypical rhythm patterns in pathological speech [5]. The speech segment, $x(t)$, is first filtered into 7 octave bands with center frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. Let $h_i(t)$ denote the filter associated with the ith octave. The filtered signal $x_i(t)$ is then denoted by,

$$x_i(t) = h_i(t) * x(t). \tag{1}$$

The envelope in the ith octave, denoted by $env_i(t)$, is extracted by:

$$env_i(t) = h_{\text{LPF}}(t) * \mathcal{H}\{x(t)\} \tag{2}$$

where, $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform and $h_{\text{LPF}}(t)$ is the impulse response of a 20 Hz low-pass filter. Once the amplitude envelope of the signal is obtained, the low-frequency variation in the amplitude levels of the signal can be examined. Fourier analysis is used to quantify the temporal regularities of the signal. With this, six EMS metrics are computed from the resulting envelope spectrum for each of the 7 octave bands, $x_i(t)$, and the full signal, $x(t)$: 1) peak frequency; 2) peak amplitude; 3) energy in the spectrum from 3-6 Hz; 4) energy in spectrum from 0-4 Hz; 5) energy in spectrum from 4-10 Hz; and 6) energy ratio between 0-4 Hz band and 4-10 Hz band. This results in a 48-dimensional feature vector denoted by $\mathbf{f}_{\text{EMS}}$.

LTAS - The long-term average spectrum (LTAS) features capture atypical average spectral information in the signal [15]. Nasality, breathiness, and atypical loudness variation, all of which are common causes of intelligibility deficits in pathological speech, present themselves as atypical distributions of energy across the spectrum; LTAS attempts to measure these cues in each octave. For each of the 7 octave bands, $x_i(t)$, and the original signal, $x(t)$, the LTAS features set consists of the: 1) average normalized RMS energy; 2) RMS energy standard deviation; 3) RMS energy range; and 4) pairwise variability of RMS energy between ensuing 20 ms frames. This results in a 28-dimensional feature vector, denoted by $\mathbf{f}_{\text{LTAS}}$.

P.563 - The ITU-T P.563 standard for blind speech quality assessment [16] is designed to measure speech quality using a parameter set that measures atypical and unnatural voice and articulatory quality. There are five major classes of features deemed appropriate for our purposes: 1) $\mathbf{f}_{\text{basic}}$ - basic speech descriptors, such as pitch and loudness information; 2) $\mathbf{f}_{\text{VT}}$ - vocal tract analysis, including statistics derived from estimates of vocal tract area based on the cascaded tube model; 3) $\mathbf{f}_{\text{stat}}$ - speech statistics, which calculate the skewness and kurtosis of the cepstral and linear prediction coefficients (LPC); 4) $\mathbf{f}_{\text{SNR}}$ - static SNR, measurements of signal-to-noise ratio, estimates of background noise, and estimates of spectral clarity based on a harmonic-to-noise ratio; and 5) $\mathbf{f}_{\text{segSNR}}$ - segmental SNR, or dynamic noise, where the SNR is calculated on a frame-by-frame basis. In the standard, a subjective rating (MOS, or Mean Opinion Score), is obtained through a non-linear combination of the above features. Here, we make use of the same feature set for our analysis, by combining all feature sets into one vector, $\mathbf{f}_{\text{VCL}}$. For a detailed description of each feature, including the mathematical derivation, please refer to [12, 16].

2.2. Cost Function Derivation

We aim to select acoustic features that discriminate well between dysarthric individuals, but that provide minimal separation between healthy speakers, as this is indicative of sensitivity to speech pathology and not speaking style. More specifically, given a complete set of features $(F) \in R^D$, we aim to select the optimal subset Ω of cardinality k, such that the learning error between the dysarthric speakers in a speaker ID task is minimized and the error between healthy speakers in the same task is minimally effected.

We start by considering the simple example of 2 dysarthric speakers and 2 healthy speakers, with feature matrices and binary label vectors given by $(\mathbf{X}_D, \mathbf{y}_D)$ and $(\mathbf{X}_C, \mathbf{y}_C)$ respectively. We model the posterior probability of a dysarthric speaker with the sigmoid $h_\theta(\mathbf{x}_D) = \frac{1}{1+e^{-\theta^T \mathbf{x}_D}}$:

$$P(y_D = 1|\mathbf{x}_D; \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}_D}} \tag{3}$$

We can write this more succinctly:

$$P(y_D|\mathbf{x}_D; \theta) = (h_\theta(\mathbf{x}_D))^{y_D}(1 - h_\theta(\mathbf{x}_D))^{(1-y_D)} \tag{4}$$

Given N_D independent training points for the two dysarthric speakers, the log-likelihood of the parameters for the dysarthric speakers can be written as:

$$\begin{aligned} l_D(\theta|\mathbf{X}_D, \mathbf{y}_D) = \quad & \sum_{i=1}^{N_D} \mathbf{y}_D^{(i)} \log h_\theta(\mathbf{X}_D^{(i)}) \\ & + (1 - \mathbf{y}_D^{(i)}) \log (1 - h_\theta(\mathbf{X}_D^{(i)})) \end{aligned} \tag{5}$$

Algorithm 1 Greedy Algorithm for Feature Selection

Input: Features and labels for dysarthric and healthy speakers: $\mathbf{X}_D, \mathbf{y}_D, \mathbf{X}_C, \mathbf{y}_C$
Output: Top k features that optimize criteria in (7): Ω
Define: $f(\Omega) = \frac{\max_\theta \, l_D(\theta|\mathbf{X}_D(\Omega),\mathbf{y}_D)}{\max_\zeta \, l_C(\zeta|\mathbf{X}_C(\Omega),\mathbf{y}_C)}$
$\Omega = \emptyset$
$F = 1 \dots M$
for $j \in 1 \dots k$ **do**
  $J = \emptyset$
  for $F_i \in F \setminus \Omega_j$ **do**
    $J(F_i) = f(\Omega \cup F_i)$
  end for
  $\Omega = \Omega \cup \{\underset{F_i}{\textbf{argmax}}\, J(F_i)\}$
end for

Similarly, we model the log-likelihood of the healthy speaker set:

$$l_C(\zeta|\mathbf{X}_C, \mathbf{y}_C) = \sum_{i=1}^{N_C} \mathbf{y}_C^{(i)} \log h_\zeta(\mathbf{X}_C^{(i)}) + (1 - \mathbf{y}_C^{(i)}) \log\left(1 - h_\zeta(\mathbf{X}_C^{(i)})\right) \tag{6}$$

Our goal is to find the subset of features that maximize the maximum likelihood of the dysarthric feature set, while simultaneously minimizing the maximum likelihood of the healthy speaker set. This results in features that accurately model the dysarthric speakers by selecting features that are specific to the disorder rather than speaking style. We formulate the following cost function:

$$\begin{aligned} \max_{\Omega} \quad & \frac{\max_\theta \, l_D(\theta|\mathbf{X}_D(\Omega), \mathbf{y}_D)}{\max_\zeta \, l_C(\zeta|\mathbf{X}_C(\Omega), \mathbf{y}_C)} \\ \text{s. t.} \quad & \text{card}(\Omega) = k. \end{aligned} \tag{7}$$

In simpler terms, the cost function aims to find the subset of features that provide minimal logistic regression error for the dysarthric feature set, while maximizing the logistic regression error for the healthy set. This results in features that provide good classification performance for the dysarthric speakers by focusing on features that do not provide good classification performance on the healthy speaker set. Although the analysis here is provided for the case of 2 speakers, we can easily extend this framework to multinomial logistic regression [17]. In fact, in Section 3, we demonstrate this framework on a set of 53 dysarthric and 13 healthy speakers.

2.3. Feature Selection Algorithm

The cost function in Eq. (7) is a good model for the task at hand; however it is difficult to solve since it optimizes over subsets of features. We approximate it using the greedy algorithm in Alg. 1.

We aim to select the top k features, denoted by Ω, from the complete set of features, denoted by F, that maximize the criteria in Eq. (7). Using a greedy approach we iteratively select the features that provide the greatest increase in the cost function. More specifically, at iteration i, we update the optimal subset using the following criteria:

$$\Omega = \Omega \cup \{\underset{F_i}{\text{argmax}}\, f(\Omega \cup F_i)\}, \tag{8}$$

where $f(\Omega)$ denotes the cost function in Eq. (7) and $F_i \in F \setminus \Omega$. A more detailed implementation is shown in Algorithm 1.

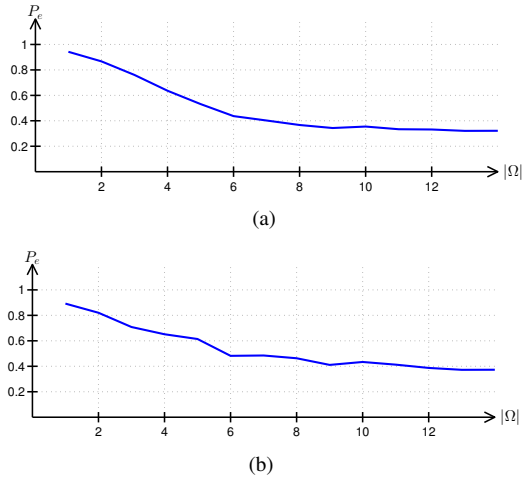

Fig. 3. Error Probability for speaker ID using features selected using Alg. 1 for (a) the dysarthric speaker set (b) the healthy set

3. RESULTS AND DISCUSSION

The feature selection algorithm described in Section 2 was implemented in MATLAB, and was used to select the top 12 features from the features in Section 2.1. A set of 34 dysarthric speakers and 13 healthy speakers were used in the study. Multinomial logistic regression was used to model the data for the dysarthric and healthy speakers [17]. The speakers were selected from a pool, collected for a larger study conducted in the Motor Speech Disorders Laboratory at Arizona State University. The dysarthria speakers included: 12 speakers with ataxic dysarthria, secondary to cerebellar degeneration, 10 mixed flaccid-spastic dysarthria, secondary to ALS, 8 speakers with hypokinetic dysarthria secondary to idiopathic Parkinson's Disease, and 4 speakers with hyperkinetic dysarthria secondary to Huntington's disease. Each speaker provided speech samples, including a reading passage, phrases, sentences, and conversational speech. While varied, the speech collection session resulted in approximately 10 minutes of recorded material per speaker, with a sampling rate of 16 kHz. The material was split into individual sentences and the features in Section 2.1 were extracted at the sentence level.

The minimum cross-validation error for the dysarthric set and the control set were used as proxies for log-likelihoods in (7). Fig 3 shows the multinomial logistic regression error for the dysarthric (34 speakers) and healthy (13 speakers) speaker feature sets after each iteration of feature selection. As shown in Fig. 3, the error for both sets drops as more features are selected. Although the final error values seem to indicate a similar error rate for both the dysarthric (37% error) and healthy speakers (40% error), it is important to note that the dysarthric set contains a total of 34 speakers, whereas the healthy speaker set contains only 13 speakers.

Table 1 shows the 12 selected features using the proposed algorithm as well as 12 features selected when using logistic regression for speaker identification (e.g. maximizing only the numerator in (7)). Here we provide a brief description of the selected features, for a more detailed discussion on computation of these features see [12, 16] and [5]. The EMS and LTAS features are extracted at different sub bands as well as for the whole signal, hence the numeric subscript in each feature's name corresponds to the center frequency of the octave band. The EMS features that were selected consist of: *Ratio40* the energy ratio between 0-4 Hz band and 4-10

Table 1. Selected Features

Speaker ID Features	Disorder Fingerprinting (Proposed Approach)
Ratio40_all	Ratio40_8000
nsd125	PV8000
Above40_all	PV2000
fPitchAverage	fSpecLevelDev
Peak_Amp_125	sd1000
Below40_125	Above40_all
E3_6Hz_125	Ratio40_all
Peak_Freq_125	Peak_Freq_125
Ratio40_125	E3_6Hz_125
Peak_Amp_250	Peak_Amp_125
Peak_Freq_250	Below40_125
Above40_125	Above40_125

Hz band; *Peak_Amp* and *Peak_Freq* the peak frequency and corresponding peak amplitude in a particular sub band; *Above40* and *Below40* the amount of energy in the spectrum below 4 Hz and above 4 Hz; and *E3_6* the energy in 3-6 Hz band. The LTAS features that were selected consist of: *PV* the pairwise variability of RMS energy between ensuing 20 ms frames; *nsd* and *sd* the normalized and unnormalized RMS energy standard deviation. The P.563 features that were selected consist of: *fPitchAverage* the average pitch; *fSpecLevelDev* the standard deviation in the band between 1-2 kHz.

Utilizing the proposed algorithm, the features that are most useful in minimizing speaker-specific discrimination, while maximizing disorder-specific discrimination tap features of rate, rhythm, and motor control patterns (i.e. top ten features identified in the analysis). For instance, features of EMS in the octave band centered around 125 Hz, including the peak frequency, amplitude of the peak frequency, and energy between 3 and 6 kHz, were all identified as important for completing this task. This is not surprising as this range is well correlated with speaking rate. Similarly, the energy above 4 kHz and the ratio of the energy above and below 4 kHz for the whole speech signal offers another representation of the rate of the speech signal. The ratio of energy above and below 4 kHz in the octave band centered around 8000 Hz was previously noted as a variable responsible for global distinction of dysarthria type [5], and offers an insight into the global distinction of the different types of speech patterns, related to the speech patterns, allowing for more idiosyncratic differences to be revealed through the features described above and below.

The LTAS and P.563 features offer information related to the fine- motor control patterns of the speakers. Features of LTAS, including pairwise variability indices in the 2000 and 8000 Hz octave bands quantify what is believed to represent articulatory precision and imprecision, respectively. These measurements quantify the change, or consistency, of energy present in 20-ms windows of the speech signal; therefore, we would expect changes in energy to offer distinct representation of a given phoneme (i.e. precise articulation). Interestingly, the P.563 feature most important for successful speaker identification was SpecLevelDev. This measurement quantified the variability of the speech signal between 1000 and 3000 Hz. Given the similarity of the pairwise variability indices, and complementary nature of the frequency range, this measurement, too, is proposed to be related to the articulatory precision with which the speech sample was spoken. Compared to the traditional speaker identification algorithms, the features identified by the proposed algorithm utilize more pathology, disorder-specific features. The traditional speaker identification algorithm utilizes more speaker-specific features. While there is some overlap in the features selected by both, there are key differences.For instance, the Speaker ID regression algorithm selects average pitch early on in the selection process, suggesting association with specific speakers. Alternatively, the P.563 and LTAS features selected with the proposed algorithm are sensitive to articulatory precision and imprecision, reflective of the motor productions specific to speech pathology.

4. CONCLUSION

In this paper, we propose a novel algorithm for feature selection that minimizes the effects of speaker-specific features and maximize the effects of disorder-specific features. The selected features therefore represent a digital signature of an individual's speech pathology state which can then be tracked over time to assess the efficacy of the provided treatment, or to sensitively track speech changes resulting from disease progression. Furthermore, we test the algorithm on a set of 34 dysarthric and 13 healthy speakers. Utilizing the proposed algorithm, selected features mostly include those corresponding to rate, rhythm, and motor control. The features selected correspond with those that are most perceptually salient in motor speech disorders, yet not isolable with a single acoustic metric. The combination of features utilized in the present investigation offers a complementary, non-redundant representation of the disrupted aspects of the speech signal. Given the instability and unreliability of subjective assessment, an objective measure of this nature is critical in the development of a gold standard for care.

5. REFERENCES

[1] J. Liss, M. Spitzer, J. Caviness, and C. Adler, "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," *The Journal of the Acoustical Society of America*, vol. 112, pp. 3022 – 3030, 2002.

[2] S. Borri and M. McAuliffe and J. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies," *Journal of Speech, Language & Hearing Research*.

[3] M. McHenry, "An Exploration of Listener Variability in Intelligibility Judgments," *Am J Speech Lang Pathol*, vol. 20, no. 2, pp. 119–123, 2011.

[4] C. Sheard, R. Adams, and P. Davis, "Reliability and Agreement of Ratings of Ataxic Dysarthric Speech Samples With Varying Intelligibility," *J Speech Hear Res*, vol. 34, no. 2, pp. 285–293, 1991.

[5] J. Liss, S. LeGendre, and A. Lotto, "Discriminating dysarthria type from envelope modulation spectra.," *Journal of Speech Language and Hearing Research*, vol. 53, no. 5, pp. 1246–55, 2010.

[6] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Using the cellular mobile telephone network to remotely monitor parkinsons disease symptom severity," *IEEE Transactions on Biomedical Engineering*, (submitted) 2012.

[7] T. Falk and W. Chan and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622 – 631, 2012.

[8] M. De Bodt and M. Huici and P. Van De Heyning, "Intelligibility as a Linear Combination of Dimensions in Dysarthric Speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, May-Jun.

[9] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, may 1982, vol. 7, pp. 1278 – 1281.

[10] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, may 1998, vol. 1, pp. 541 –544 vol.1.

[11] S. Voran, "Objective estimation of perceived speech quality. II. Evaluation of the measuring normalizing block technique," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 4, pp. 383 –390, Jul. 1999.

[12] "Single-sided speech quality measure," ITU-T Recommendation P.563, International Telecommunication Union, Telecommunication Standardization Sector, 2004.

[13] S. B. Davis and P. Mermelstein, "Evaluation of acoustic parameters for monosyllabic word identification," *Journal of the Acoustical Society of America*, vol. 64, no. S1, pp. S180–S181, 1978.

[14] R. Vergin and D. O'Shaughnessy, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 525–532, Sept. 1999.

[15] P. Rose, *Forensic Speaker Identification*, Taylor and Francis, London.

[16] L. Malfait, J. Berger, and M. Kastner, "P.563 - The ITU-T Standard for Single-Ended Speech Quality Assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1924 –1934, Nov. 2006.

[17] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, New York.