Spectral Features for Automatic Blind Intelligibility Estimation of Spastic Dysarthric Speech

Richard Hummel, Wai-Yip Chan

Queen's University
Dept. of Electrical and Computer Eng.
Kingston, Ontario, Canada

Tiago H. Falk

Institut National de la Recherche Scientifique
Energy, Materials, and Telecommunications
Montréal, Québec, Canada

Abstract

In this paper, we explore the use of the standard ITU-T P.563 speech quality estimation algorithm for automatic assessment of dysarthric speech intelligibility. A linear mapping consisting of three salient P.563 internal features is proposed and shown to accurately estimate spastic dysarthric speech intelligibility. Delta-energy features are further proposed in order to characterize the atypical spectral dynamics and limited vowel space observed with spastic dysarthria. Experiments using the publicly-available Universal Access database (10 speaker patients) show that when salient delta-energy and internal P.563 features are used, correlations with subjective intelligibility ratings as high as 0.98 can be attained.

Index Terms: Dysarthria, intelligibility, P.563, subjective quality, delta-energy

1. Introduction

Patients with cerebral palsy may develop a speech disorder, called "spastic dysarthria", which can severely affect the intelligibility of their speech [1]. The symptoms of spastic dysarthria include distorted vowels and consonants, reduced speaking rate, harsh voice quality, and abnormal prosody (e.g., uniform loudness and pitch) [1]. The treatment of speech disorders usually includes assessment of intelligibility in order to gain a better picture of the patient's condition. The most common method of assessing intelligibility is through a listener based (subjective) assessment. Drawbacks of subjective assessment include the inherent variability of human listener ratings and the cost of a clinician's time. Familiarity with the dysarthric speaker's particulars may bias the intelligibility estimate [2]; furthermore, listeners may weigh perceptual dimensions differently to estimate intelligibility, resulting in inconsistencies. By using a computer based approach, an objective, repeatable method can be crafted.

In the effort to create such a computer based method, there have been a few different approaches. Researchers have developed automatic speech recognition (ASR) systems, trained on healthy speech, that use the recognition rate as a measure of intelligibility [2][3] This approach assumes that ASR systems will behave similarly to human perception when presented with dysarthric speech. An alternate approach involves training the intelligibility estimation system with dysarthric speech [4]. This approach, however, still relies on an ASR system trained with normal speech.

ASR systems can be difficult to train, and usually require a template 'reference' signal or knowledge about the phonetic content of the target word. To estimate intelligibility without vocabulary restriction, features which do not require a reference signal are desired. These features are known as 'blind' or 'single-ended'. Many systems exist for computing blind estimates of subjective telephone speech *quality*, but it is not known if such systems can be modified to estimate *intelligibility* of dysarthric speech. While speech quality and intelligibility are related, their relationship is not trivial [5]. For example, broad bandwidth speech may be intelligible and pleasant, whereas synthesized speech may be intelligible but artificial sounding and therefore deemed poor in quality.

In this paper, we first investigate existing blind assessment methods for evaluating speech quality. To that end, we explore the "ITU-T P.563" [6] standard, which is a method for blind, objective speech quality estimation. The P.563 standard describes an algorithm for determining the dominant type of distortion of the input speech signal, and then combines relevant features based on the chosen class to compute a mean opinion score (MOS) estimate. After examining the P.563 standard, we propose new blind features better suited to estimating intelligibility in dysarthric subjects. The proposed features are delta-energy coefficients derived from a mel-spaced triangular filterbank. The use of delta instead of absolute energy values reduces speaker dependence and better reflects the uniform loudness typical of spastic dysarthria. Subbands close to the extreme ranges of the F1-F2 vowel space are shown to be the most correlated with intelligibility. Next, we update the feature mapping defined in P.563 to improve correlation with subjective intelligibility scores. Re-mapping P.563's features for a particular purpose has precedence; for example, the work in [7] proposed a new feature mapping to improve MOS estimates

of noise suppressed speech. Finally, we show that further improvement can be attained if delta-energy features are included in the revised mapping.

2. ITU P.563: Internal features and an updated mapping function

As previously discussed, P.563 is a standard for the blind estimation of narrow-band speech quality, namely, the mean opinion score (MOS). Before estimating the MOS value, P.563 first classifies the distortion present in the input signal into one of six 'distortion classes'. The six classes cover distortions related to noise, clipping, muting, and 'unnatural' voice. Class assignment is performed using a hierarchical approach that chooses the most relevant distortion class when multiple distortion types are detected. The reasoning behind this is psycho-acoustic; some types of distortion affect subjective quality scores more than others. Based on the selected class, P.563 selectively combines some of its 43 internal features to compute a final MOS estimate. The 43 features include parameters related to vocal tract shape, signal noise characteristics and linear predictive coding (LPC) parameters. For more details, refer to [6].

In this paper, we examine the MOS estimates from the six distortion classes, as well as the 43 internal features. Despite not being originally designed for our purpose, one or more of the parameters may show significant correlation with intelligibility. Furthermore, we propose a new mapping which linearly combines salient internal P.563 features to estimate intelligibility. The P.563 features included in the model are to be chosen based on how well they rank in an iterative feature selection process.

Let ϕ_{est} be the intelligibility estimate, N the number of features, $\mathcal{F}$ the set of selected features, and a_i the model coefficient for feature f_i. Our proposed model can then be written as

$$\phi_{est} = a_0 + \sum_{i=1}^{N} a_i f_i, \quad f_i \in \mathcal{F}. \tag{1}$$

3. Proposed Features: Mel-frequency delta-energy coefficients

Several spectral subbands are critical for speech intelligibility; prior work has shown that intelligibility is mostly preserved when only sparse spectral information is kept [8]. Furthermore, energy values in particular subbands are likely to be sensitive to the imprecise consonants and distorted vowels typical of spastic dysarthria; subbands lying on the extremities of the F1-F2 space should be particularly sensitive due to a phenomenon known as 'vowel centralization' [9]. Energy content above 4000 Hz should also decrease in dysarthric speech [10].

Center frequencies of the triangular filters are linearly spaced on the mel scale, and range in value from 89 to 7016 Hz. Utterances are first normalized before feature extraction to -26 dBov by the ITU-P56 voltmeter [11]. A discrete Fourier transform and a shifting 50 ms Hamming window with 50% overlap are computed over speech segments to get the magnitude spectrum. The filterbank energy values are

$$M_{fc}(i) = \sum_{k=0}^{L/2} |X(i,k)|^2 Y_{fc}(k), \tag{2}$$

where $X(i,k)$ denotes the k^{th} Fourier transform coefficient of the i^{th} windowed frame, L is the length of the Fourier transform, fc is the center frequency of a triangular filter, and $Y_{fc}(k)$ the triangular filter coefficients. Let N_w be the number of frames. The delta mel-band energies are finally defined as:

$$\delta_{fc} = \frac{1}{N_w - 1} \sum_{i=1}^{N_w - 1} |M_{fc}(i+1) - M_{fc}(i)|. \tag{3}$$

4. Experimental results

4.1. Universal access database

The data used in this paper is a subset of the Universal Access database [12], which contains single word recordings of cerebral palsied dysarthric subjects. We use the 10 speakers with spastic dysarthria that have corresponding subjective intelligibility scores. The vocabulary represented in the database consists of 300 uncommon English words, and 155 other words. The 300 uncommon words were selected from children's novels, and the 155 other words consist of spoken single digits (10 words), words from the radio alphabet (26 words), computer commands (19 words), and the 100 most common English words from the Brown Corpus of Written English. Each speaker has 765 files in the datbase, consisting of one utterance each of the 300 'uncommon' words, and three utterances of each of the 155 'other' words.

To assess intelligibility, a subset of 200 words in total was taken from all word categories; 25 of the 200 words were repeated to assess intra-speaker reliability. Five naive judges, all native speakers of American English between 18 and 40 years of age, were employed to transcribe what they heard for each word. The five transcription accuracy percentages were averaged to give one intelligibility score per speaker. We label these average scores as $\{\phi_1, \phi_2, .., \phi_{10}\} = \Phi$. They are used as our 'ground truth' intelligibility ratings, with which we measure the performance of our proposed estimators.

4.2. P.563 internal features and updated mapping

4.2.1. P.563 MOS estimates

The 765 utterances per speaker were first downsampled to 8 kHz, and then processed by P.563. For each of the six distortion classes, an average MOS value is calculated over each speaker's 765 utterances; Pearson's correlation

coefficient (denoted r) is then computed between the averages and Φ. The result is that none of the six classes of MOS estimates are significanty correlated ($p < 0.05$) with Φ. Therefore, we disregard the MOS estimates as potential intelligibility estimators and focus on designing a new mapping better suited to our task.

4.2.2. Feature selection

To select a suitable set of features for the new mapping, we run M trials of a feature selection process which selects and ranks N features every trial. Following the M trials, average feature rank (AFR) is employed to select a final set of features for our mapping.

Each feature selection trial proceeds as follows. First, 233 of the 465 word vocabulary are randomly selected (including 150 uncommon words). Utterances corresponding to the selected vocabulary serve as data for feature selection and model training. Feature selection begins by discarding any feature not significantly correlated with Φ ($p > 0.05$); sequential feature selection (SFS) [13] then selects N of the remaining features, with r serving as selection criterion. The sequentially selected features are given increasing rank values from one to N, with one being the highest rank. The selected features constitute a temporary feature set $\mathcal{F}_i$ used to form the linear model for trial i, with model coefficients obtained using least squares regression. For model validation, unseen data (i.e., not used for feature selection and model training) is applied to the trained model to compute one ϕ_{est} value per speaker. Model performance is quantified using r, Spearman's rank correlation (ρ), and root mean square error (ϵ) between ϕ_{est} values and Φ.

Feature selection is repeated M times for a given value of N, we average the M values of r. ρ and ϵ and denote the results $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ to give measures of average model performance. To determine which features were selected most often, we use average feature rank (AFR) to obtain a final feature list of length N. AFR can be expressed as

$$AFR_i = \frac{1}{M} \sum_{j=1}^{M} R_i(j), \tag{4}$$

where $R_i(j)$ is the rank of feature f_i for trial j. Since each trial selects N features, if feature f_i is not selected for trial j, $R_i(j) = N + 1$. The features with the N smallest AFR values then become our final choice for $\mathcal{F}$.

4.2.3. Updated P.563 feature mapping

The first task in designing a new P.563 feature mapping is finding a suitable value for N, the number of included features. To do this, we vary N from one to 20 and examine $\bar{r}$ as N increases; for each value of N, we compute $\bar{r}$ using $M = 200$ feature selection trials. The 43 P.563 features serve as the candidate feature pool. Performance begins to degrade severely for $N > 6$; therefore, performance results versus number of features, up to $N = 6$, are plotted in Figure 1. We choose $N = 3$ as it offers near optimal performance with a small number of features. Average performance values $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ for $N = 3$ are listed in Table 1.

Estimator	Performance Measure		
	$\bar{r}$	$\bar{\rho}$	$\bar{\epsilon}$ (%)
P.563 re-mapping	0.95	0.93	11.0
Composite	0.98	0.95	7.1

Table 1: *Mean performance measures from P.563 and composite estimator experiments*

The high average correlation values (up to 0.95) show that linearly combining internal P.563 features is a better approach to estimating intelligibility than using P.563's MOS estimates. Now that N has been chosen, AFR can be applied to the results from the $M = 200$ feature selection trials. The top three selected features, ordered by AFR, are listed in Table 2. To better understand why the listed features are useful, each one is discussed in turn.

Feature Pool	AFR	Feature Name
P.563	1.00	LPCcurt
	2.70	CepCurt
	3.34	SpecLvlRange
P.563 + proposed	1.00	LPCcurt
	2.04	δ_{992}
	3.03	δ_{1361}

Table 2: *AFR values of top ranked features*

The top two ranked features, LPCcurt (LPC kurtosis) and CepCurt (cepstum kurtosis), characterise abnormal deviations in the sample kurtosis values of LPC and cepstral coefficients. Kurtosis characterises the 'peakedness' of a distribution. Normal speech produces LPCcurt and CepCurt values within certain ranges; P.563 maps deviations from expected ranges to low MOS scores.

The third highest ranked feature, SpecLvlRange (spectral level range), is calculated by computing the average difference between the 85^{th} and 20^{th} percentiles of the spectral amplitude distribution (estimated from a short term Fourier transform). The relevance of this feature lies in the hypernasal nature of spastic dysarthric speech. Many dysarthric speech samples in the database have one prominent low frequency formant, causing the average difference between the two percentile values to increase.

4.3. Proposed features

Before incorporating proposed features into our feature selection process, we first survey their behavior with respect to intelligibility. Average feature values are computed over each speaker's 765 utterances; then, r and ρ are calculated between average feature variables and Φ, as was done with the MOS estimates. Twelve of the

20 δ_{fc} features show significant ($p < 0.05$) r correlation with Φ and are listed in Table 3.

The properties of dysarthric speech corroborate the results presented here. All listed correlations are *positive*, relating to the uniform loudness voice typical of spastic dysarthria. Coefficients from subbands with center frequencies around 1000 and 2500 Hz correlate strongly with Φ. These frequencies lie roughly on the extremeties of the F1-F2 space. Table 3 shows that delta energy values for subbands above 3500 Hz are also strongly correlated with Φ, corroborating our hypothesis regarding high frequency energy.

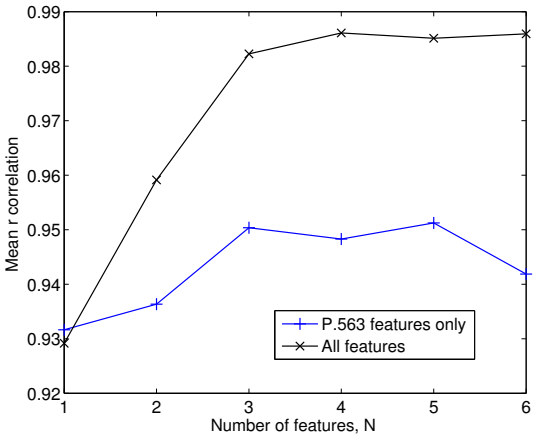

Figure 1: *Estimator performance as a function of N*

4.4. Composite estimator

The final experiment involves combing proposed features with P.563 features into a composite feature pool and performing feature selection as before. The experimental procedure from Section 4.2.3 is re-used. Again, $N = 3$ is chosen based on optimizing $\bar{r}$ as N varies from one to six (Figure 1). As can be seen, incorporating proposed features improves results for $N > 1$. Mean performance measures and top ranked features are presented alongside previous results in Tables 1 and 2, respectively. Features δ_{992} and δ_{1361}, displace SpecLvlRange and CepCurt from the previous experiment. AFR values for the top three features are lower than those obtained previously. Thus, features are top ranked in more trials than before. Interestingly, features δ_{992} and δ_{1361} are not the highest ranked in Table 3. Closer inspection reveals that δ_{992} and δ_{1361} aided in the estimation of intelligibility scores for moderate-mild dysarthria (i.e., greater than 50%).

5. Conclusion

We have examined the capabilities of P.563, a standard for blind estimation of mean opinion scores (MOS), and determined that a new mapping of internal P.563 features was able to outperform MOS estimates as predictors of intelligibility on dysarthric speech. Additional blind features were proposed, and a composite estimator was designed that incorporates proposed with internal P.563 features to achieve further improvements. Results indicate that automatic assessment using blind features is a promising method for estimating dysartric speech intelligibility.

Feature	r	ρ
δ_{4075}	0.84	0.82
δ_{1128}	0.83	0.85
δ_{4683}	0.81	0.75
δ_{3535}	0.80	0.82
δ_{2631}	0.76	0.78
δ_{992}	0.76	0.81
δ_{738}	0.75	0.76
δ_{1361}	0.73	0.73
δ_{5367}	0.72	0.65
δ_{3056}	0.70	0.81
δ_{6144}	0.68	0.65
δ_{1624}	0.63	0.65

Table 3: δ_{fc} *features significantly correlated with* Φ

6. References

[1] D. Freed, *Motor speech disorders: Diagnosis and treatment*. Singular Pub Group, 2000.

[2] J. Carmichael and P. Green, "Revisiting dysarthria assessment intelligibility metrics," in *Proc. Intl. Conf. Spoken Lang. Proc.*, 2004, pp. 742–745.

[3] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Noth, "PEAKS-A system for the automatic evaluation of voice and speech disorders," *Speech Commun.*, vol. 51, no. 5, pp. 425–437, 2009.

[4] C. Middag, G. Van Nuffelen, J. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in *Proc. Intl. Conf. Spoken Lang. Proc.*, 2008, pp. 1745–1748.

[5] J. Deller Jr, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1993.

[6] ITU-T P.563, "Single-ended method for objective speech quality assessment in narrowband telephony applications," Intl. Telecom. Union, 2004, geneva, Switzerland.

[7] A. Ekman and W. Kleijn, "Improving quality prediction accuracy of P. 563 for noise suppression," in *Proc. Intl. Workshop Acoustic Echo and Noise Control*, 2008.

[8] S. Greenberg, T. Arai, and R. Silipo, "Speech intelligibility derived from exceedingly sparse spectral information," in *Proc. Int. Conf. Speech and Lang. Proc.*, 1998, pp. 2803–2806.

[9] H. Liu, F. Tsao, and P. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *J. Acoust. Soc. Am.*, vol. 117, pp. 3879–3889, 2005.

[10] R. Kent, G. Weismer, J. Kent, H. Vorperian, and J. Duffy, "Acoustic Studies of Dysarthric Speech: Methods, Progress, and Potential." *J. Commun. Disord.*, vol. 32, no. 3, pp. 141–86, 1999.

[11] ITU-T P.563, "Objective measurement of active speech level," Intl. Telecom. Union, 1993, geneva, Switzerland.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Intl. Conf. Spoken Lang. Proc.*, 2008, pp. 1741–1744.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn Res.*, vol. 3, pp. 1157–1182, 2003.