QUANTIFYING PERTURBATIONS IN TEMPORAL DYNAMICS FOR AUTOMATED ASSESSMENT OF SPASTIC DYSARTHRIC SPEECH INTELLIGIBILITY

Tiago H. Falk

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

Richard Hummel, Wai-Yip Chan

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

ABSTRACT

Spastic dysarthric speech is often associated with imprecise placement of articulators which, in turn, cause perturbations in speech temporal dynamics, such as unclear distinctions between adjacent phonemes. While these perturbations can lead to a significant reduction in intelligibility, measures to objectively assess their detrimental effect on intelligibility are lacking. In this paper, short- and long-term temporal dynamics measures are proposed and evaluated as correlates of subjective intelligibility. The former is based on log-energy temporal dynamics information, whereas the latter is based on an auditory-inspired modulation spectral signal representation. A composite measure is also developed based on linearly combining the proposed measures with a tone-unit duration parameter. Experiments with the publicly-available 'Universal Access' database of spastic dysarthric speech show that the proposed composite measure can achieve rank correlations with subjective ratings as high as 0.87, thus providing a tool to automatically diagnose speech disorder severity and to evaluate dysarthria treatment outcomes.

Index Terms— Dysarthria, temporal dynamics, intelligibility, modulation spectrum, cepstrum.

1. INTRODUCTION

Dysarthria is a motor speech disorder, resultant from damage to the central and/or peripheral nervous systems, which affects articulation, speech rates, and prosody, amongst other intelligibility-reducing factors (e.g., vocal harshness). The most prominent subtype of dysarthria is commonly associated with cerebral palsy and is called "spastic dysarthria" [1]. Symptoms of spastic dysarthric speech can include imprecise placement of articulators, atypical speech rates, incomplete consonant closure, and monotonicity [1]. Today, therapists mainly depend on subjective intelligibility tests to characterize speech disorder severity and to evaluate treatments.

Subjective tests, however, are costly, laborious, and subject to many listener biases (e.g., familiarity with the patient's speech disorder). Objective measurement methods, on the other hand, replace the listener panel with a computational algorithm, thus can offer an economical and reliable alternative to subjective assessment. Objective measures aim to deliver scores that are highly correlated with the intelligibility ratings obtained from subjective listening experiments. In fact, there is growing evidence that objective measures are being used to assist clinicians in dysarthria treatment decisions [2].

Objective methods can be classified as either *blind* or *comparison-based*. Comparison-based methods depend on a reference signal or feature prototype of a target word being uttered. In [3], for example, the Itakura-Saito distortion was computed between the produced disordered speech utterance and the same utterance spoken by a healthy individual. In order to account for differences in utterance durations, dynamic time warping was applied. Alternately, small-vocabulary automatic speech recognition has become a popular method of intelligibility assessment for speakers with mild or moderate dysarthria; technological advances, however, are still needed before ASR is used for severe dysarthric speakers [4]. An additional limitation of ASR is the sparseness of available data needed to accurately train speaker-dependent acoustic models which have been shown to outperform those obtained from "healthy" natural speech [5].

In many practical applications reference signals and/or features may not be available and blind measures are needed. Today, the majority of existing blind measures depend on quantifying atypical prosody and parameters such as fundamental frequency (*f0*) variation and second-formant slope transitions [6]. Recent studies, however, have shown that articulation errors are the major contributing factor to reduced intelligibility in dysarthric speech [7], followed by prosody, voice quality, and nasality. Representative symptoms of imprecise placement of articulators can include prolonged phonemes, unclear distinction between adjacent phonemes, odd speech rates, and rhythmic disturbances, to name a few [1]. Hence, it is expected that the development of blind measures of speech temporal dynamics perturbations will allow for accurate dysarthric intelligibility estimation. In this paper, several measures are described and evaluated on the publicly-available 'Universal Access' dysarthric speech database.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

2. QUANTIFYING PERTURBATIONS IN SPEECH TEMPORAL DYNAMICS

Several factors are known to adversely affect speech intelligibility for individuals with dysarthria, the most prominent being odd temporal dynamics, and disordered prosody [7]. Here, short-term and long-term temporal dynamics measures are explored as indicators of dysarthric speech intelligibility. An additional parameter, namely, word duration within tone units, is also explored as it captures temporal distortions affecting prosody. These measures, along with several benchmark ones, are described in more detail below.

2.1. Short-term Temporal Dynamics

The zeroth order cepstral coefficient is computed as a measure of short-term log-spectral energy and the zeroth order delta coefficient is used as a measure of log-energy rate of change. Let $c_0(m)$ denote the zeroth order cepstral coefficient for frame m. $\Delta c_0(m)$ represents the zeroth order delta coefficient and is computed as

$$\Delta c_0(m) = \sum_{l=-L}^{L} l \, c_0(m+l), \tag{1}$$

where the normalization factor $\sum_{l=-L}^{L} l^2$ is omitted as it does not affect the results. In our simulations $L = 3$ is used and cepstral coefficients are computed over 32-millisecond frames with 10-millisecond frame shifts. In order to characterize short-term temporal dynamics oddity, the standard deviation (σ_Δ) of the N Δc_0 samples within a speech file is used

$$\sigma_\Delta \quad = \quad \sqrt{\frac{1}{N-1}\sum_{m=1}^{N}(\Delta c_0(m) - \bar{\Delta} c_0)^2}, \tag{2}$$

where $\bar{\Delta} c_0$ indicates the sample average of $\Delta c_0(m)$.

2.2. Long-Term Temporal Dynamics

Long-term speech temporal dynamics information is captured by means of an auditory-inspired modulation spectral signal representation which characterizes the rate of change of long-term speech temporal envelopes. To obtain the modulation spectral representation, the dysarthric speech signal is first filtered by a bank of 23 gammatone critical-band filters, which emulate the processing performed by the cochlea. The filter center frequencies range from 50 Hz to approximately half the sample rate; filter bandwidths are characterized by the equivalent rectangular bandwidth. Long-term temporal dynamics information is captured from the Hilbert temporal envelope found for each of the 23 gammatone filter outputs.

Hilbert temporal envelopes are windowed and a discrete Fourier transform is used to compute the modulation spectrum for each frame. Lastly, modulation frequency bins are grouped into eight bands in order to emulate an auditory-inspired modulation filterbank [8]. Second-order bandpass filters with a quality factor $Q = 2$ are used and modulation filter center frequencies range from $2 - 64$ Hz. The k^{th} modulation band energy for frame m is denoted as $E_{j,k}(m)$, $k = 1, \ldots, 8$. Additionally, the modulation energy averaged over N speech frames is given by

$$\bar{E}_{j,k} = \frac{1}{N}\sum_{i=1}^{N} E_{j,k}(i). \tag{3}$$

It is known that healthy natural speech contains dominant modulation frequencies from $2 - 20$ Hz with spectral peaks at approximately 4 Hz [9]. It is hypothesized that prolonged phonemes, slower speech rates, as well as the unclear distinction between adjacent phonemes caused by imprecise placement of articulators, will cause a shift of the modulation frequency content to modulation frequencies below 4 Hz. In turn, as intelligibility levels increase, modulation frequency content will be better spread across higher modulation frequencies, as observed with natural speech [10]. In order to characterize this oddity in speech temporal dynamics, the ratio of modulation energy at frequencies below 4 Hz to modulation frequencies above 4 Hz is proposed. The low-to-high modulation energy ratio (LHMR) is given by

$$\text{LHMR} = \frac{\displaystyle\sum_{k=1}^{K^*}\sum_{j=1}^{23}\bar{E}_{j,k}}{\displaystyle\sum_{k=K^*+1}^{8}\sum_{j=1}^{23}\bar{E}_{j,k}}, \tag{4}$$

where K^* corresponds to the index of the modulation filter centered at approximately 4 Hz; in our experiments, $K^* = 4$.

2.3. Voicing Duration

An additional temporal parameter is computed, namely, the total duration of voiced segments within the uttered word. The measure is represented by $\mathcal{V}$ and has been used in the past to characterize speech disorders [11].

2.4. Benchmark Measures: "EMS"

In [12], several temporal envelope modulation spectral (EMS) features were proposed and used to characterize long-term dynamics in dysarthric speech. In the mentioned study, the modulation spectrum was computed in a manner different than that described in Section 2.2. More specifically, the dysarthric speech was first filtered through a bank of seven octave-band filters with center frequencies ranging from $125 - 8000$ Hz. Temporal envelopes were then extracted from the full-band signal as well as from the seven sub-band signals via half-rectification and lowpass filtering at 30 Hz. The envelopes were downsampled to 80 Hz, mean-subtracted and the modulation spectrum was computed (up to 10 Hz) via a 512-point FFT using a Tukey window.

A number of features were then extracted from the computed modulation spectrum and six salient features were used to estimate dysarthric speech intelligibility. The features were B_{250} and B_{8000} (normalized EMS energy below 4 Hz); A_{4000} (normalized EMS energy between 4-10 Hz); R_{2000} (ratio of B_{2000} to A_{2000}); E_{250} and E_{4000} (normalized EMS energy between 3-6 Hz), where the subscript indicates the center frequency of the octave-band filter used in the computation of the measure. These features are used as benchmark measures.

It is important to emphasize, however, that due to limitations in our dataset (sampled at 16 kHz, as opposed to 44.1 kHz as in [12]) calculation of the measure B_{8000} was not possible. To ameliorate this shortcoming, a slightly different filterbank was used in our simulations. In lieu of 7 octave-band filters, 8 critical-band gammatone filters were used with center frequencies ranging from 125-5200 Hz and bandwidths from approximately 80-600 Hz. With this modification, the parameter R_{2000} was computed as the average of the ratio parameter computed for the fifth and sixth critical-band filters, centered at 1.5 kHz and 2.4 kHz, respectively. Similarly, parameters E_{4000} and A_{4000} were computed by averaging the respective parameters over the seventh and eighth filters, centered at 3.4 kHz and 5.2 kHz, respectively. Lastly, the measure B_{8000} was computed from the last filter centered at 5.2 kHz. Despite these modifications, the aforementioned notation is used throughout the remainder of this paper.

2.5. Composite Measures

To test the complementariness of the developed and benchmark measures, two composite measures are also explored

$$\begin{aligned} f_{proposed} &= \alpha_0 + \alpha_1 * \sigma_\Delta + \alpha_2 * \text{LHMR} + \alpha_3 * \mathcal{V}, \\ f_{EMS} &= \beta_0 + \beta_1 * R_{2000} + \beta_2 * B_{8000} + \beta_3 * B_{250} \\ &+ \beta_4 * E_{4000} + \beta_5 * E_{250} + \beta_6 * A_{4000}. \end{aligned}$$

To estimate the weights α_i and β_i, the available data is partitioned into disjoint training and test sets (see Section 3.2) and least-squares linear regression is used.

3. EXPERIMENTAL RESULTS

3.1. Database: UA-Speech

The data used in our experiments consisted of the audio content of the Universal Access (UA-Speech) audio-visual database [13]. Speech data from ten participants diagnosed with spastic dysarthria due to cerebral palsy were used in our experiments; participant demographics is shown in Table 1. Each participant read a total of 765 isolated words displayed on a computer screen, spread over three blocks of 255 words, including 155 words that were repeated in each block and 100 uncommon words that differed across blocks. The repeated words consisted of the 10 digits, 26 radio alphabet letters, 19 computer commands, and the 100 most common words in the Brown corpus of written English. The 300 uncommon words (100 per block) were selected from children's novels [13].

Table 1. Demographics of the ten spastic dysarthric speakers

Subject	Gender	Age	Intelligibility	Category
1	male	18	2%	very low
2	male	18	15%	very low
3	male	58	28%	low
4	male	unreported	43%	low
5	male	21	58%	mid
6	male	40	91%	high
7	male	28	93%	high
8	female	51	6%	very low
9	female	30	29%	low
10	female	22	95%	high

For the subjective intelligibility test, five listeners were recruited per speaker. Listeners were between the ages of 18-40, native speakers of American English, had no prior experience with disordered speech, and had no previous training in phonetic transcription. Listeners were instructed to provide orthographic transcriptions of 225 speech utterances presented via headphones in a quiet environment. The 225 utterances consisted of 10 digits, 25 radio alphabet letters, 19 computer commands, and 73 words randomly selected from each of the common and uncommon word categories, plus 25 arbitrarily chosen words that were repeated in order to assess intra-listener reliability, which remained around 92%. Listener transcriptions were then analyzed and the mean percentage of correct responses, averaged across the five listeners, was calculated to obtain the subjective intelligibility score of each dysarthric speaker.

3.2. Results

Table 2 reports the Pearson (R) and Spearman rank (R_S) correlation coefficients, along with their corresponding $p-$values, for the developed and benchmark measures. R_S is used to quantify how similar the measures rank with subjective listening ratings. As can be seen, all proposed measures achieved significant Pearson and Spearman correlations ($p < 0.05$) with subjective intelligibility ratings. On the other hand, only two of the benchmark measures achieved significant Pearson correlations, namely B_{250}, E_{250}; only the latter also achieved a significant Spearman correlation.

In order to obtain the weights of the composite measures, the UA-Speech database was partitioned into two disjoint sets. Speech files belonging to the 'uncommon word' category (300 files per participant) served as (unseen) test data and the remaining files (465 files per participant) served as training data. As can be seen from Table 2, the composite measure based on the three proposed measures ($f_{proposed}$)

Table 2. Performance of proposed and benchmark measures. Performances reported for composite measures $f_{proposed}$ and f_{EMS} are based on the unseen test set comprised of speech files belonging to the 'uncommon word' category.

Measure	R	p	R_S	p
σ_Δ	0.76	0.01	0.80	0.01
LHMR	-0.75	0.01	-0.67	0.03
$\mathcal{V}$	-0.79	0.01	-0.73	0.02
$f_{proposed}$	0.85	0.01	0.87	0.01
R_{2000}	-0.61	0.06	-0.41	0.25
B_{8000}	0.40	0.25	0.22	0.50
B_{250}	0.64	0.05	0.62	0.06
E_{4000}	-0.30	0.40	-0.33	0.35
E_{250}	0.70	0.02	0.72	0.02
A_{4000}	-0.30	0.40	-0.28	0.40
f_{EMS}	0.79	0.01	0.69	0.03

achieved $R = 0.85$ and $R_S = 0.87$ on the unseen test set. On the other hand, the benchmark composite measure based on the six EMS features (f_{EMS}) achieved $R = 0.79$ and $R_S = 0.69$. Hence, $f_{proposed}$ attains an approximate 29% correlation-improvement ($R\%$) relative to f_{EMS} for Pearson correlation and a 58% correlation-improvement in Spearman correlation. $R\%$ is computed as

$$R\% = \frac{R_{proposed} - R_{EMS}}{1 - R_{EMS}} \times 100\%, \qquad (5)$$

and reflects the percentage reduction of the EMS-based composite measure's performance gap to perfect correlation.

3.3. Discussion

The proposed LHMR measure proposed here differs from the benchmark R_{2000} measure in several manners. First, LHMR incorporates information from 23 acoustic frequency bands and not just the band centered at 2 kHz. Second, the use of the Hilbert transform for temporal envelope calculation allows for modulation frequencies beyond 10 Hz to be incorporated; such higher frequencies are important for intelligibility estimation [10]. Lastly, the proposed LHMR measure is based on an auditory-inspired modulation filterbank which is used to group modulation frequency bins according to psychoacoustic insights. With the R_{2000} measure, simple averaging of Fourier transform-derived frequency bins is performed. Moreover, our implementation of the benchmark measures differed slightly from those reported in [12]. A direct comparison between the results reported there, however, suggests that the modifications did not affect performance (e.g., $R = 0.69$ was reported for parameter R_{2000}). The differences obtained in correlation are likely due to utterance types and not the filterbank modifications; in [12], sentences were used, whereas with the UA-Speech database, single words were uttered.

4. CONCLUSION

In this paper, three measures were described to characterize temporal dynamics perturbations and shown to correlate significantly with subjective intelligibility ratings and to outperform several benchmark metrics. A composite measure was also developed and shown to be a reliable indicator of dysarthric speech intelligibility, thus providing a means to automatically evaluate dysarthria treatments.

5. REFERENCES

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby, 2005.

[2] A. Hill et al, "An Internet-based telerehabilitation system for the assessment of motor speech disorders: a pilot study," *American Journal of Speech-Language Pathology*, vol. 15, no. 1, p. 45, 2006.

[3] L. Gu, J. Harris, R. Shrivastav, and C. Sapienza, "Disordered speech assessment using automatic methods based on quantitative measures," *EURASIP Journal Applied Signal Proc.*, vol. 9, pp. 1400–1409, 2005.

[4] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features," *EURASIP Journal Advances Signal Proc.*, 2009, Article ID: 629030, 9 pages.

[5] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[6] M. Klopfenstein, "Interaction between prosody and intelligibility," *International Journal of Speech-Language Pathology*, vol. 11, no. 4, pp. 326–331, 2009.

[7] M. De Bodt, M. Hernández-Díaz 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, 2002.

[8] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I – model structure," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[9] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Intl. Conf. Speech and Lang. Proc.*, Oct. 1996, pp. 2490–2493.

[10] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[11] 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 Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[12] S. LeGendre, J. Liss, and A. Lotto, "Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra," *Journal of the Acoustical Society of America*, vol. 125, p. 2530, 2009.

[13] 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.