TOWARDS A CLINICAL TOOL FOR AUTOMATIC INTELLIGIBILITY ASSESSMENT

Visar Berisha, Rene Utianski, and Julie Liss

Department of Speech and Hearing Science, Arizona State University

Abstract

An important, yet under-explored, problem in speech processing is the automatic assessment of intelligibility for pathological speech. In practice, intelligibility assessment is often done through subjective tests administered by speech pathologists; however research has shown that these tests are inconsistent, costly, and exhibit poor reliability. Although some automatic methods for intelligibility assessment for telecommunications exist, research specific to pathological speech has been limited. Here, we propose an algorithm that captures important multi-scale perceptual cues shown to correlate well with intelligibility. Nonlinear classifiers are trained at each time scale and a final intelligibility decision is made using ensemble learning methods from machine learning. Preliminary results indicate a marked improvement in intelligibility assessment over published baseline results.

Index Terms*:* intelligibility assessment, speech pathology, machine learning, multi-scale analysis

1. INTRODUCTION

Speech production is an intricate process of motor coordination, requiring muscle groups from many subsystems, including respiration, phonation, resonance, and articulation. In healthy individuals, the planning and implementation of these motor movements is fluently and accurately executed in time and space, resulting in a clear speech output. It follows that disruption to any of the aforementioned subsystems required for speech will manifest as degradations at different time scales in the speech output. Assessing the severity of intelligibility reduction due to speech output degradation has been a long-standing problem in the field. Most often this is done by subjective assessment by an expert in the area.

Subjective tests are inherently inconsistent, costly and, oftentimes, not repeatable. In fact, research has shown poor inter- and intra-rater reliability in clinical assessment [1, 2]. Furthermore, clinicians working with clients form a bias based on their interactions, resulting in unreliable intelligibility assessment [2]. As such, a set of valid, reliable, objective, and *sensitive* metrics of speech intelligibility are desired.

Unfortunately, procedures that would enhance validity and reliability of subjective evaluation (e.g., sizeable listener panel, inclusion of anchor conditions, etc. [3, 4, 5]) are costly and time consuming. Over the past several decades, research has begun to capitalize upon computer-based evaluations, with the goal to offer a repeatable, reliable assessment with minimal cost. Two major approaches are utilized: 1) reference-based intelligibility estimations, which measure deviation from a "clean" reference signal, and 2) blind assessments, which measure a variety of speech features, irrelevant of the intended target.

A number of reference-based approaches rely on estimating subjective intelligibility through the use of pre-trained automatic speech recognition (ASR) algorithms [6]. More specifically, these algorithms are trained on healthy speech and the error rate on pathological speech serves as a proxy for estimating the intelligibility decrement [6].

In addition to ASR-based approaches, other reference-based approaches quantify perceptual differences between the distorted signal and an "oracle" clean signal [7, 8, 9, 10, 11]. The metrics rely on simplified psychoacoustic models that mimic human perception to measure perceptual errors. Although these algorithms have been shown to correlate well with subjective assessment, their utility in assessing pathological speech is limited because of the requirement for a reference signal. In contrast to these reference-based approaches, we propose an algorithm that operates only on the degraded speech and generates an estimate of intelligibility.

Research in blind algorithms for intelligibility assessment has been more limited. In telecommunications, the ITU-P.563 standard has been shown to correlate well with speech quality, however this is not optimized for pathological speech [12]. In [13], [14] and [15], the authors attempt to estimate dysarthric speech intelligibility using a set of selected acoustic features. Although the algorithms have shown some success in a narrow context, the feature sets used in these papers do not make use of long-term rhythm disturbances in the signal, common in the dysarthrias. In contrast to this, the algorithm we propose here makes use of acoustic cues at different time scales that capture short-term voicing problems and long-term rhythm problems.

Motivated by previous research on perceptual correlates to intelligibility and quality [16,12], we propose a system that extracts features at different time scales (resolution of phonetic, segmental, and suprasegmental information) and attempts to assess specific challenges to intelligibility. As mentioned above, the algorithm does not use a reference signal and it relies on features extracted at multiple scales. Generally speaking, these features measure distorted rate and timing of speech (sentence level features), unnatural loudness variation (sentence level features), unnatural pitch/formant variation (vowel/consonant level features), articulatory imprecision (vowel/consonant level features), and omissions or distortions of specific consonants and vowels (phoneme level features). See Table 1 for examples of proposed metrics for each level of analysis. For each scale, we train a classifier, and, as a result, obtain multiple intelligibility decisions per sentence. A final decision is made using ensemble learning methods from machine learning. The classifiers are trained on a training set and tested on a development set. All results are presented for the "NKI CCRT Speech Corpus" (NCSC) recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute .

The rest of this paper is outlined as follows: In section 2, we detail the technical approach, including a description of the features at different scales and the classification algorithms used. In section 3, we provide comparative results using the pathology challenge data. In section 4, we close with some concluding remarks and outline future work.

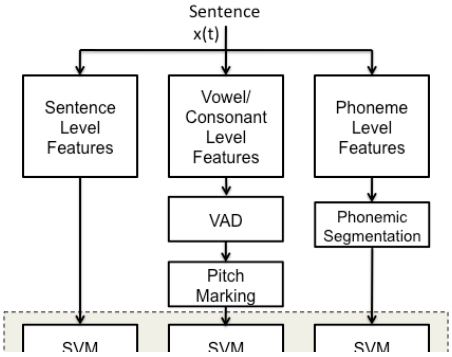

Figure 1: A high-level diagram of the proposed classification framework. Features are extracted at three different scales, and a final intelligibility decision is made by combining individual decisions from each scale.

2. TECHNICAL APPROACH

A high-level diagram of the proposed approach is shown in Fig. 1. As the diagram shows, for each speech sentence, $x(t)$, we extract a set of features at three different scales: at the sentence level, at the vowel/consonant level, and at the phoneme level. Pre-trained SVMs are used to make an intelligibility decision at each level, then a linear stacking scheme fuses the individual classifier decisions to form a final hypothesis. In this section, we describe in detail the main contributions of this paper: the novel feature set and the classification scheme used to make final intelligibility decisions.

2.1. Features

2.1.1 Sentence Level Features

Baseline – The 6125-dimensional baseline feature set in [17] was used as a starting point. In an effort to manage the data size and to minimize the effects of the curse of dimensionality, we use PCA to reduce the dimension of this feature space by retaining the PCA features that account for 90% of data energy (389 features). We denote the resulting feature set by $\mathbf{f}_{\text{bl}}$.

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. It has been shown to be a useful indicator of atypical rhythm patterns in pathological speech [16]. 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 i^{th} octave. The filtered signal $x_i(t)$ is then denoted by,

$$x_i(t) = h_i(t) * x(t) \quad (1)$$

The envelope in the i^{th} octave, denoted by $e_i(t)$, is extracted by:

$$e_i(t) = h_{\text{LPF}}(t) * H(x(t)) \quad (2)$$

Level	Feature set
Sentence	
Rate, rhythm, and prosody	EMS
Nasality, breathiness, loudness variation	LTAS
Vowel/ Consonant	
Unnatural pitch/formant contours	Basic speech descriptors
Vowel space reduction	Formant structure statistics
Articulatory Imprecision	Vocal tract statistics
Vocal quality (nasality and breathiness)	Spectral energy distribution
Phoneme	
Distortions/substitutions	Silence statistics, spectral statistics
Rate and rhythm	Duration

Table 2: Perceptual cues previously shown to correlate to decreased speech intelligibility and their respective feature set proxies.

where, $H(.)$ is 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. Six EMS metrics were then 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) captures atypical average spectral information in the signal. Nasality, breathiness, and atypical loudness variation, 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 full signal, $x(t)$ we extract 1) Average normalized RMS energy, and 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 $\mathbf{f}_{\text{LTAS}}$.

Combining all three sentence-level feature sets into one vector, we obtain the following 465-dimensional feature vector:

$$\mathbf{f}_{\text{SL}} = \left[\mathbf{f}_{\text{bl}}^{\text{T}} \quad \mathbf{f}_{\text{EMS}}^{\text{T}} \quad \mathbf{f}_{\text{LTAS}}^{\text{T}}\right]^{\text{T}}$$

This becomes the input of our sentence level SVM classifier.

2.1.2 Vowel/Consonant Level Features

A voice-activity detector (VAD), followed by a voicing detector is used to locate all active voiced speech frames. The VAD operates on 4*ms* frames and is based on an iterative adaptive thresholding approach, similar to that in [12]. The voicing detector, followed by an autocorrelation-based pitch estimator on 64*ms* frames is used to determine pitch marks, for the pitch-synchronous features described below. For the specifics of the

pitch marking algorithm, we refer the reader to [12]. This pre-processing is required for extracting the salient features for intelligibility assessment. For all active frames, we calculate some basic descriptors of the speech signal and some more advanced speech statistics. For all voiced frames, we track the vocal tract statistics. We describe the features in detail below.

Basic Speech Descriptors – For all active speech frames, we calculate a series of features that describe the basic properties of the speech signal. More specifically, we calculate 1) the average pitch, 2) an estimate of pitch variance, 3) the average signal level, and 4) the signal level variance for active frames, and 5) the signal level variance for voiced frames. The resulting 5-dimensional vector is denoted as $\mathbf{f}_{\text{basic}}$.

Speech Formant Statistics – It has been shown that the kurtosis and skewness of the speech formant structure serve as good cues for unnatural speech [12, 17]. For each active speech frame, we calculate the 21st-order LPC and cepstral coefficients. Then for each set of coefficients, we calculate the per-frame kurtosis (κ) and skewness (ζ) of the LPC and cepstral coefficients, using (3) and (4) below:

$$\kappa = \frac{1}{\mathrm{P}}\sum_{i=1}^{P}\left(\frac{\mathrm{a_i} - \frac{1}{P}\sum_{k=1}^{P}\mathrm{a_k}}{\sigma_{\text{coeff}}}\right)^4 \quad (3)$$

$$\zeta = \frac{1}{\mathrm{P}}\sum_{i=1}^{P}\left(\frac{\mathrm{a_i} - \frac{1}{P}\sum_{k=1}^{P}\mathrm{a_k}}{\sigma_{\text{coeff}}}\right)^3 \quad (4)$$

where P is the order of the formant analysis and σ_{coeff} is the standard deviation of the coefficients. The average and standard deviation of these values over the entire active speech signal serve as the salient speech statistics for unnatural speech detection. The resulting 4-dimensional vector is denoted by $\mathbf{f}_{\text{stat}}$.

Vocal Tract Statistics – In an attempt to directly estimate physical properties of the speaker vocal tract, we model the vocal tract as a set of tubes of time-varying cross sectional area [12]. The area is estimated using the reflection coefficients from pitch-synchronous windows. Pitch-synchronous analysis allows for windows that are synchronized with the human speech production system. We calculate the 8th order reflection coefficients from the LPC coefficients for each window and then calculate the tube areas using the following equation

$$\mathrm{S_i} = \frac{1+\mu_i}{1-\mu_i}\mathrm{S_{i+1}}, \mathrm{i} = 8, 7, \ldots, 1 \quad (5)$$

The tube areas are then combined into a set of features representing the rear (S_1, S_2, S_3), middle (S_4, S_5, S_6), and front (S_7, S_8) articulators for every voiced frame. For each speech segment, we calculate the following features 1) The maximum value of S_1 over the voiced sections of the speech signal, 2) The average value of S_8 over the voiced sections of the speech signal, 3) The averaged cross-sectional area of the rear articulators over the speech signal, 4) An estimate of the correlation between the cross-sectional area of the rear and middle articulators 5) The average of the derivative of the position of the maximum cross-sectional area of all eight tubes (this measures the consistency with which the tube changes over time), 6) The ratio of voiced frames in the speech signal over all active speech frames. The resulting 6-dimensional vector is denoted by $\mathbf{f}_{\text{VT}}$.

Combining all three feature sets into one vector, we obtain the following 15-dimensional feature vector:

$$\mathbf{f}_{\text{VCL}} = \left[\mathbf{f}_{\text{basic}}^{\text{T}} \quad \mathbf{f}_{\text{stat}}^{\text{T}} \quad \mathbf{f}_{\text{VT}}^{\text{T}}\right]^{\text{T}}$$

This becomes the input of our vowel/consonant level SVM classifier.

2.1.3 Phonemic Level Features

The phonemic information file for the data in [17] is used to segment the sentences into their composite phonemic parts. For each phoneme in a sentence, we extract the features described below. The features become inputs to different classifiers, each trained on specific phonemes, and an intelligibility decision is formed for each phoneme present in a given sentence. The features are described below.

Phoneme Duration – The duration of each phoneme is readily available from the phonemic information file. The duration is indicative of rhythmic problems in pathological speech. The duration is denoted by $\mathrm{f_{PD}}$.

Phoneme Bandwidth – The average bandwidth for each phoneme in the sentence is estimated by computing the frequency range corresponding to 80% of all signal energy. We denote this feature by $\mathrm{f_{PBW}}$.

Silence Statistics – We calculate the length of the silences and their normalized position in the speech signal. More specifically, for each sentence, we calculate the average duration of each silence and the average normalized position. The duration is readily available from the phonemic information file. The normalized position is calculated by dividing the midpoint of each silence segment by the total duration of the speech signal. The resulting 2-dimensional vector $\mathbf{f}_{\text{sil}}$ is used to

Combining all feature sets into one vector, we obtain the following 4-dimensional feature vector:

$$\mathbf{f}_{\text{PL}} = \left[\mathrm{f_{PD}} \quad \mathrm{f_{PBW}} \quad \mathbf{f}_{\text{sil}}^{\text{T}}\right]^{\text{T}}$$

This becomes the input of our phoneme level SVM classifier.

2.2. Training

Prior to training, all classifier inputs are standardized to mean 0 and standard deviation 1. The standardization parameters are learned from the training set and applied to the cross-validation/test set during learning. For all trained SVMs, optimal parameter selection was done through cross-validation. More specifically, we select initial values of the complexity parameter and the RBF kernel parameter from the following ranges respectively: $C \in 2^{[-5:2:15]}$ and $\gamma \in 2^{[-15:2:3]}$; this is then followed by a more refined search centered around the optimal value. The LIBSVM toolbox for Matlab was used for SVM classification [18].

2.3. Ensemble Learning

Motivated by its success in the recent Netflix competition [19], we make use of linear classifier stacking as a way of combining the scores from multiple classifiers. The base-level classifiers are individually trained on the training data, each resulting in a different probability of success (PS) on the development set. We use the normalized PS as weights for our linear combination scheme, followed by a thresholding operation. Note that not all classifiers are present for each sentence since each sentence contains different phonemes.

2.4. Results

All results are presented for the "NKI CCRT Speech Corpus" (NCSC) recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute [17]. The corpus contains recordings and intelligibility evaluations of 55 speakers, resulting in 1647 labeled sentences and 739 unlabeled sentences in the test set. The labeled data is further divided into a training set and a development set. Although a partition was provided for analysis in [17], we noticed significant differences in data statistics between these two sets. This could be due to different recording conditions, different speakers, different phrases, etc. As a result, we select a random subset of examples to serve as a new training set and a new development set. The statistics of the resulting partitions were much more consistent when compared to the originals.

In Table 2, we show the unweighted and weighted average (UA/WA) recall of the intermediate SVMs (at the sentence and at the vowel/consonant level) and of the final ensemble method. These results are compared against the baseline classification results published in [17]. As the table shows, the proposed approach results in a marked improvement in recall rate for the development set.

Classifier	Development UA (WA)
Baseline – SVM	61.4 (61.3)
Baseline – RF	65.1 (65.1)
SVM – Sentence	79.6 (79.8)
SVM – VC	76.6 (77.3)
Ensemble	**84.4 (84.8)**

Table 2: Recall rates for the proposed classification schemes compared against the baseline results.

3. CONCLUSIONS

The ultimate goal of this work is to develop robust methods for predicting listener performance with a given pathological speech signal. Toward that end, future analyses will explore the utility of the proposed approach to predict different dimensions of intelligibility decrements. Different measurements of "intelligibility," such as identifying words, word onsets, and phonemes, that track directly to the acoustic measurements at the sentence, vowel/consonant, and phoneme level may show a more robust relationship to how this manifests in a listener's percept.

Further, examining this relationship will allow for the determination of when "degraded" speech becomes problematic for listeners, offering a quantifiable assessment of the severity of the communication impairment. By examining aspects of production, more fine- grained measures of perceptual features (i.e., subjective descriptions vs. binary measures of "unintelligible" or "intelligible"), in tandem with relative acoustic measurements, we may be able to better understand the process from production to perception.

4. References

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

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

[3] K.M. Yorkston, K.M. "Treatment Efficacy: Dysarthria," in Journal of Speech & Hearing Research, Volume 39, Issue 5, pp. S46- 57, 1996.

[4] K.M. Yorkston, et al. "Evidence for Effectiveness of treatment of loudness, pitch, or prosody in dysarthria: A
systematic review," in Journal of Medical Speech-Language Pathology, Volume 12, Issue 2 , pp. xi-xxxvi, 2007.

[5]C. Sellars, T. Hughes, P. Langhorne. "Speech and language therapy for dysarthria due to non-progressive brain damage," in Cochrane Database of Systematic Reviews 2005, Issue 3.

[6] P. Doyle, H. Leeper, A. Kotler, N. Thomas-Stonell, C. O'Neill, M. Dylke, and K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34.

[7] V. Berisha and A. Spanias, "Wideband speech recovery using psychoacoustic criteria," *EURASIP Journal on Audio, Speech, and Music Processing*, Volume 2007 Issue 2, April 2007.

[8] D.H. Klatt, "Prediction of perceived phonetic distance from critical band spectra: a first step", in *Proc of IEEE ICASSP*, 1278-1281, 1981.

[9] W. Yang, M. Benbouchta, and R. Yantorno, "A modified bark spectral distortion measure as an objective speech quality measure", in *Proc. of IEEE ICASSP*, 541-544, 1998.

[10] S. Voran, "Estimation of perceived speech quality using measuring normalizing blocks", in *Proc. of IEEE Spch. Cod. Wksp.*, 83-84, 1997.

[11] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs" ITU-T, ITU-T Rec. P.862, 2001.

[12] "Single-sided speech quality measure" ITU-T, ITU-T Rec. P.563, 2004.

[13] T. Falk, W. Chan, F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility" *Speech Communication*, ScienceDirect, 2011.

[14] M. De Bodt, H.M Hernandez-Diaz, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, 2002.

[15] R. Hummel, "Objective Estimation of Dysarthric Speech Intelligibility", *Thesis, Masters of Applied Science,* Queen's University Kingston, Ontario, Canada Sept. 2011.

[16] J.M. Liss, S. Legendre, and A.J. Lotto. "Discriminating dysarthria type from envelope modulation spectra," in *Journal of Speech, Language, and Hearing Research,* Volume 53, pp. 1246-1255, 2010.

[17] JY Lee, J. Sangbae, and H. Minsoo, "Automatic Assessment of Pathological Voice Quality Using Higher-Order Statistics in the LPC Residual Domain." *EURASIP Journal on Advances in Signal Processing*, 2010.

[18] C.C. Chang and C.J. Lin, "LIBSVM : a library for support vector machines." in *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011.

[19] J. Sill et al. "Feature Weighted Linear Stacking" *arXiv:0911.0460*, 2009.