A Screening Test for Speech Pathology Assessment Using Objective Quality Measures

Eric J. Wallen and *John H.L. Hansen**
Robust Speech Processing Laboratory
Department of Electrical Engineering
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech

ABSTRACT

Vocal dysfunctions and pathologies can be devastating to one's ability to produce speech properly. A novel method for approaching the problem of speech pathology assessment is presented in this paper. The focus is not to detect or measure all possible pathologies, but rather to assess quality for the case where the probablility of pathology is high. The system is a screening test that combines objective quality measures that examine both excitation and vocal tract characteristics. The five integrated measures are pitch perturbation, amplitude perturbation, a main cepstral peak measure, the log-likelihood measure, and an energy-weighted log-likelihood measure. They are evaluated over six speech phoneme classes and their ability to assess the quality of speech is examined. Ultimately, these measures will be seamlessly integrated into an overall pathology assessment system using a hidden Markov model (HMM) recognizer. To demonstrate the ability of the quality measures to probe the multidimensional perceptual quality space, a neural network based speech pathology detection scheme was established. This system attained an average classification rate of 85.8% for healthy and pathology speech.

1. INTRODUCTION

The need for a reliable method of screening subjects for speech production abnormalities is becoming apparent in today's society. Legislation was recently passed which requires all preschool children to be tested for speech and hearing disorders prior to entering school [1, 2]. Traditionally, assessment of speech quality has been conducted subjectively by trained speech pathologists [3, 4]. However, given the task of screening a large number of subjects, such a method can be extremely inefficient. In addition, given a case where it is determined that the subject is producing low quality speech, it may be necessary to probe further with physically invasive techniques such as a laryngoscopy. Applying such invasive techniques add to the inefficiency of this method in terms of time and cost and are also uncomfortable for the subject. Hence, the need is evident for a fast and noninvasive automated method of screening for speech abnormalities. Again, the focus here is not to distinguish between speech pathologies, but to concentrate on the assessment of speech quality as an indicator of more potentially serious problems in speech production.

Signal processing techniques offer many advantages over traditional pathology assessment methods. They provide the prospect of early detection of speech disorders which would lead to a greater possibility for recovery. Secondly, such techniques allow for a quantitative and noninvasive procedure for the measurement of vocal quality. Finally, they have the promise of being able to screen larger patient populations so that speakers at risk are promptly referred to a speech pathologist. It is clear that the application of such a screening test would be vastly more efficient, both in terms of time and cost, than traditional approaches.

Although traditional methods of speech pathology detection are fairly reliable, they are time consuming and expensive. Their ability to detect abnormalities in speech production in their early developmental stages is less likely since patients sometimes wait until quality is more seriously affected. Historically, most objective speech quality measures have been formulated for assessing the performance of speech vocoders [5]. More recently, several measures have been considered for assessment of severe pathology, such as vocal fold cancer [6]. This paper will describe a novel approach for integrating objective measures of speech quality with a proposed hidden Markov model (HMM) detection scheme for the purpose of developing a screening test capable of assessing subjects who are not producing high quality speech, a condition that may be indicative of a more serious speech disorder.

2. ALGORITHM DEVELOPMENT

In the process of formulating an algorithm to address this problem, it is necessary to determine what characteristics would typify both neutral/healthy and abnormal/unhealthy speech. Having established these conditions, the focus then shifts to developing a screening system that incorporates features that are able to appreciably measure these characteristic differences. The objective quality measures that are incorporated into the proposed screening test were chosen to take advantage of certain inherent differences between speech that is produced under neutral and healthy conditions and speech that is subject to abnormalities caused by some

*This work was sponsored by a biomedical research grant from The Whitaker Foundation.

physiological or neurological impairment. It is precisely this difference which provides the basis for the proposed screening system.

It has been hypothesized that one characteristic of speech produced under non-healthy conditions is an irregular periodicity. Since this is essentially a function of the excitation of the speech production system, it is necessary to incorporate objective quality measures that focus on excitation characteristics in order to successfully track such differences. Two of the five measures chosen for this system are based on this hypothesis. An additional measure utilizes excitation information in its analysis, and the remaining two measures emphasize characteristics which reflect pathology in the vocal tract and/or frequency response.

The specific measures considered include (1) pitch perturbation and (2) amplitude perturbation measures (excitation based measures), the (3) log-likelihood and (4) energy-weighted log-likelihood measures, and (5) the cepstrum measure, which offer a spectral based assessment of both excitation and vocal tract characteristics.

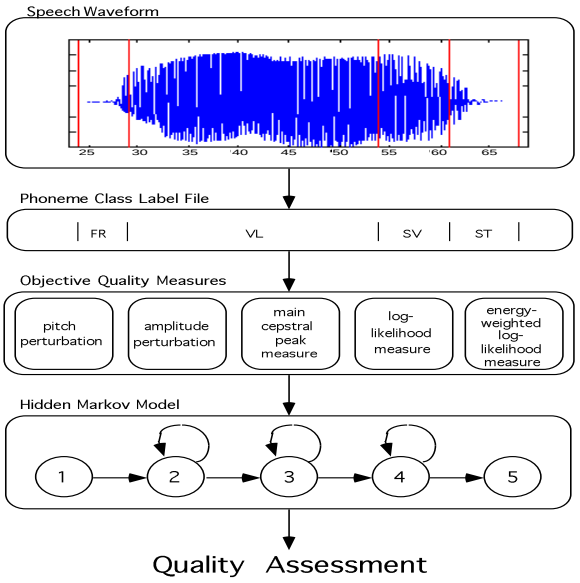

Figure 1: *Proposed pathology assessment system block diagram.*

3. PATHOLOGY ASSESSMENT SYSTEM

The diagram in Figure 1 illustrates the overall proposed pathology assessment system. The first step is the partitioning of the speech into broad phoneme classes (vowel, diphthong, semivowel, nasal, stop, fricative). This is important because certain phoneme classes are more sensitive to pathology than others, and a particular objective quality measure may be better suited to track these sensitivities. The phoneme-class partitioning is achieved using a previously formulated algorithm, originally designed for text-directed speech enhancement [7]. The parser can also be used here, since the screening test assumes that speakers will be producing a particular sequence of words and sentences in response to computer based verbal or visual cues. Secondly, a dedicated feature extraction is performed for each class, and the five objective speech quality measures are applied. Finally, these measures are employed in a phoneme-class dependent manner to obtain an overall measure of the quality or probability of pathology using the proposed HMM detection scheme. The five objective measures will now be considered.

3.1 Pitch Detector:

The method of pitch estimation employed in this study was based on a scheme using the dyadic wavelet transform (D_yWT) developed in [8] and implemented in[9]. The D_yWT detects the instant of glottal closure, where the distance between two such events constitutes the pitch period for that epoch. This scheme was chosen because it was capable of estimating the widely varying pitch periods often encountered in speech pathology, especially for speakers with smaller vocal folds and vocal tracts (children and female speakers who have significantly higher mean pitch values).

3.2 Pitch Perturbation:

The pitch perturbation measure implemented here is based on the work of Feijoo and Hernández [6], which quantifies how the pitch profile changes with time. The hypothesis is that a subject suffering from a speech disorder will produce speech that is not as fluid, strong, or constant as their healthy counterpart. By analyzing the time evolution of the pitch information, an understanding of the degree to which a subject is producing smooth and fluid speech in obtained. If the extracted pitch period sequence is defined as SP, the pitch pertubation measure (PP) is written,

$$PP = (\frac{1}{N-1})(\frac{1}{SP_{max}}) \sum_{i=1}^{N-1} |SP(i+1) - SP(i)| \quad (1)$$

Pitch estimation errors can sometimes cause extraneous spikes in the pitch period profile. To address this, a 3-point median smoothing algorithm was applied to the pitch period sequence prior to the calculation of the pitch perturbation response.

3.3 Amplitude Perturbation:

The amplitude perturbation (AP) measure [6] tracks the time evolution of a set of main amplitude peaks of the time-domain signal. The amplitude feature, denoted by SA, is determined by the maximum amplitude of the signal within a moving analysis window. This measure is useful because it quantifies the mismatch between adjacent amplitude peaks, thus providing an overall measure of the smoothness of the time-domain signal. The measure is written as

$$AP = (\frac{1}{N-1})(\frac{1}{SA_{max}}) \sum_{i=1}^{N-1} |SA(i+1) - SA(i)| \quad (2)$$

3.4 Main Cepstral Peak:

The cepstrum is a measure that contains spectral information about the excitation of a voiced signal and the vocal tract characteristics. For this study, we only consider that portion of the cepstrum related to excitation, which is located in the higher quefrency region

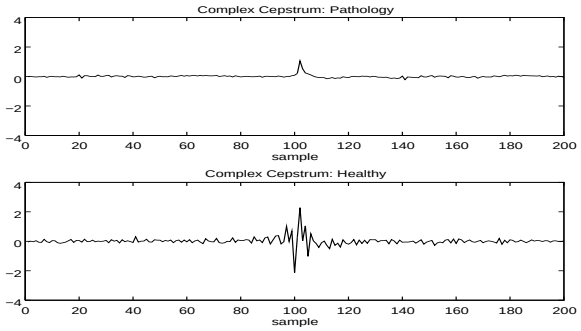

Figure 2: *High quefrency region of the complex cepstrum for (a) vocal fold cancer patient and (b) healthy subject, illustrating differences in amplitude of the main cepstral peak.*

of the cepstrum signal. While the shape of the glottal pulse sequence will also effect the overall spectral slope, and thereby the low quefrencey characteristics, we choose not to consider that trait here. The complex cepstrum is defined as

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(e^{j\omega})|e^{j\omega n}d\omega \qquad (3)$$

It has been found that a degraded periodicity of the signal is represented by a decrease in the amplitude of the main cepstral peak in the high quefrency range of the cepstrum signal. This can be clearly seen in Figure 2, which is an example of an adult vocal fold cancer patient and a healthy subject. Taking advantage of this fact, the main cepstral peak measure is able to detect when a subject is producing speech of degraded periodicity, and hence of lower quality. This measure (CE), which was explored in [6], is defined as

$$CE = \frac{1}{NS}\sum_{i=1}^{NS} PC_i \qquad (4)$$

where PC_i is the amplitude of the maximum positive cepstral peak of the ith segment and NS is the total number of segments used in the measure. For Figure 2, the pathology and healthy values are $PC_{102} = 1.15$ and $PC_{103} = 2.21$, respectively.

3.5 Log-Likelihood Measures:

The log-likelihood measure has traditionally been used in the evaluation of vocoders [5] or enhancement algorithms [10] by comparing the original (reference) waveform with the coded or enhanced waveform. In the application of speech pathology, we are not provided with a reference, so a modified version is used based on a frame-by-frame analysis. Since the measure is based on the dissimilarity between the all-pole models of adjacent frames, the result is a tool for quantifying the level of spectral mismatch within a particular segment of speech. Defining x_ϕ as the reference frame and x_d as the adjacent frame, the log likelihood measure is written as,

$$d(\overrightarrow{a_d}, \overrightarrow{a_\phi}) = log(\frac{\overrightarrow{a_d} R\phi \overrightarrow{a_d}^T}{\overrightarrow{a_\phi} R\phi \overrightarrow{a_\phi}^T}) \qquad (5)$$

where $\overrightarrow{a_\phi}$ and $\overrightarrow{a_d}$ represent the LPC coefficient vectors for the reference and adjacent frames, respectively, and $R\phi$ is the autocorrelation matrix for x_ϕ.

The energy-weighted log-likelihood measure factors into the above measure a weighting term based on the energy of the reference frame being considered. It is written as,

$$d_w(\overrightarrow{a_d}, \overrightarrow{a_\phi}) = E(x_\phi) * d(\overrightarrow{a_d}, \overrightarrow{a_\phi}) \qquad (6)$$

where $E(x_\phi)$ is the energy of the reference frame.

4. EVALUATION METHOD

The data used in this paper was obtained from a study conducted by the Dept. of Pediatrics, Duke Medical Center [11]. A total of twenty-one (eleven female and ten male) pre-school aged native speakers of English were analyzed. The data, which was supplied in VHS video format, was digitized at an 8kHz sampling rate. The basis for the screening test was two tokens each of seven isolated words and five sentences. The isolated words used in the study were *bees, spot, chick, sneeze, keys, seal*, and *rat* and the sentences are listed below.

1. *Now I'll plant the bushes and trees.*
2. *Mother says look what is here.*
3. *Some bluebirds are eating the seed.*
4. *We saw flying fish at the zoo.*
5. *They walked into the store to buy his books.*

This speech corpus represents a portion of the larger data set available on VHS video tape.

5. RESULTS

In this evaluation, we will first establish a criterion for low quality speech using the five measures across broad phoneme classes, then integrate these results into the formulation of an overall assessment method.

5.1 Quality Measure Results:

Each of the objective quality measures is computed for specific phoneme classes across the entire database of isolated words and complete sentences. The results are listed in Table 1. Measures that are not computed for a certain phoneme class are indicated with an x in the appropriate location. The issue addressed here is the analysis of the information-bearing characteristics of each phoneme class as well as the appropriateness of each quality measure in examining that class. For example, diphthongs represent a movement of the articulators, which would be indicated in the spectral characteristics of the utterance. Since the log-likelihood measure tracks spectral mismatch, it is well suited for analysis of this phoneme class. Nasals, which are generally characterized by an absence of spectral energy in certain locations, are also appropriate for the log-likelihood measures. As an obvious example of a case where a particular measure is not appropriate for a phoneme class, consider a measure where voicing is required, such as pitch perturbation. Examining an unvoiced fricative with this measure, due to its lack of periodic structure, is inappropriate. Therefore, a set of rules are established for measure application in this algorithm.

Mean Objective Quality Measures					
Phone Class	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*
Vowel	0.041	0.050	3.11	0.25	0.25
Diphthong	0.064	0.042	3.03	0.21	0.21
Semivowel	0.075	0.055	3.38	0.43	0.43
Nasal	0.062	0.061	3.14	0.35	0.35
Stop	x	0.072	3.78	0.35	0.36
Fricative	x	0.085	3.72	0.28	0.29

Table 1: Mean objective quality measures across broad phoneme classes. (a) pitch perturbation, (b) amplitude perturbation, (c) main cepstral peak measure, (d) log-likelihood measure, and (e) energy-weighted log-likelihood measure.

5.2 Quality Measure PDFs:

Experimental probability density functions (PDFs) were created from the quality measures computed across each phoneme class. Employing these distributions and an informal subjective quality listening test, some initial conclusions regarding speech quality among the twenty-one children were drawn. The subjective listening test was designed to determine how well the child was able to produce what this study deemed to be high quality speech, taking into account such attributes as intelligibility and articulation. Using the results from this listening test as a general guide, instances of relatively low and high quality speech were then compared to the experimental PDFs, which provided a connection between subjective and objective measures. Figure 3 gives an example of the PDFs for the log-likelihood measure across vowels and fricatives.

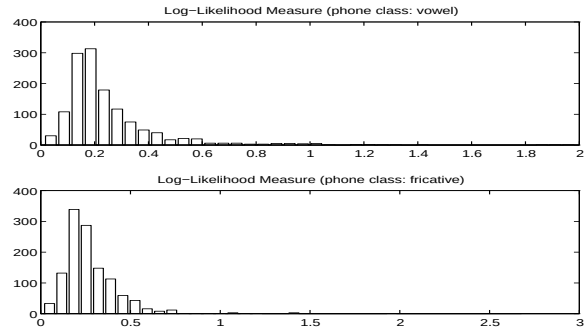

Figure 3: *Distributions of the log-likelihood measure for vowel and fricative phone classes.*

5.3 Classification Performance:

While the ultimate focus here will be the use of an HMM recognizer, we considered a classification method based on neural networks. A system was developed using the objective measures in conjunction with a neural network scheme for the detection of vocal fold cancer (VFC). The quality measures provided the input to a feed-forward network trained with the backpropagation learning algorithm. A database of twenty adult VFC patients and nine healthy subjects was used to train and test this detection scheme. Using an open round-robin training method, which consisted of twenty test trials of six speakers each, the system made only one error in classifying a cancer patient as healthy. This corresponds to a classification rate of 98.3% for this twenty-nine speaker set. The number of healthy speakers classified as pathology (false positives) was somewhat higher, resluting in an average classification rate for the healthy and cancerous patients of 85.8%.

These results are encouraging, however it is important to note that the neural network scheme might be considered to be a more general pattern classifier. It is believed that the HMM based classifier will be better able to accurately characterize the time-evolution of the objective measure sequence across the word and sentence test corpus.

6. CONCLUSIONS

The assessment of speech pathology is a difficult problem to address. In this study, a screening test that integrates a series of objective quality measures with a hidden Markov model (HMM) scheme was proposed. The focus of the screening test was to assess the subjects who were not producing speech of high quality, which we hypothesized might be indicative of a more serious speech disorder. Five objective quality measures were considered and their appropriateness across phoneme classes was examined. Although the HMM recognizer has yet to be fully developed, an average classification rate of 85.8% for healthy and pathology speakers was obtained using a neural network based approach.

References

[1] Bill # H.R. 5357, Congressional Record, June, 1992.

[2] Individuals with Disabilities Education Act (IDEA), P.L. 102-119, Washington, D.C., 1991.

[3] M. Hirano, S. Hibi, T. Yoshida, H. Kasuya, and Y. Kikuchi, "Acoustic analysis of pathological voice, some results of clinical application," *Acta Otolaryngologica*, 105:432-438, 1988.

[4] P. Kitzing, "Simultaneous photo- and electroglottographic measurements of voice strain," In I. R. Titze & R. C. Scherer, editor, *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, pages 221-229, 1983.

[5] S. Quackenbush, T. Barnwell III, M. Clements. *Objective Measures of Speech Quality*, Prentice Hall, New Jersey, 1988.

[6] S. Feijoo, C. Hernández, "Short-term stability measures for the evaluation of vocal quality," *J. Speech & Hearing Research*, 33:324-334, 1990.

[7] B. Pellom, J.H.L. Hansen, "Text-Directed Speech Enhancement Using Phoneme Classification and Feature Map Constrained Vector Quantization," IEEE Int. Conf. on Acoust., Speech, and Signal Processing, 1996, pp. 645-648.

[8] S. Kadambe, G.F. Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Trans. on Information Theory*, vol. 38, pp. 917-924, 1992.

[9] D. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," *The Journal of the Acoustical Society of America,* vol. 96, no. 6, pp. 3392-3400, December 1994.

[10] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Proc.,* 39(4):795-805, April 1991.

[11] R.A. Sturner, J. Heller, S.G. Funk, M. Feezor, "Preschool Hearing Screening via Speech Signals," *ASHA,* vol. 33, 1991.