Automatic Assessment of the Spasmodic Voice

Mark A. Bartsch

April 23, 2002

Abstract

The present work proposes automatic methods for determining the extent of dysphonia in patients with spasmodic dysphonia. Spasmodic dysphonia is a disorder of the larynx which involves involuntary contractions of the vocal folds during speech production. Numerous acoustical aspects of a voiced spasmodic speech signal which indicate the presence of these vocal spasms have been identified, including voicing breaks, shifts in fundamental frequency, and signal aperiodicities. We have developed algorithms for assessing the severity of dysphonia which are based on periodicity analysis, fundamental frequency estimation, and the processing of signals in the time-frequency and time-lag domains. These methods are tested on thirty-seven recordings of patients with spasmodic dysphonia reading a phonetically balanced passage of text. The algorithm's assessments are evaluated by comparing them to ratings generated by trained listeners on the same data set. A baseline algorithm achieves a mean rank correlation of 0.76 when compared to the five judges, as compared with a mean rank correlation of 0.83 among the judges.

1 Introduction

Spasmodic dysphonia (SD) is formally characterized as a focal dystonia of the larynx [1]. Patients affected by SD exhibit abnormal contractions of the laryngeal muscles during vocal production. The disorder produces a wide variety of effects in the speech of an afflicted patient; dysphonic speech has been described as strained-strangled, squeezed, effortful, choked, jerky, hoarse, and groaning [18]. Vocal tremor is reported to accompany SD in roughly one-quarter of cases [1]. One of the most widely used treatments for SD is the injection of Botulinum Toxin A (Botox) into the muscles around the larynx. The injections serve to paralyze the muscles affected by the abnormal contractions, providing some measure of relief from symptoms [5].

Assessment of SD, as with many vocal disorders, is dependent in large part on a perceptual evaluation of the patient's vocal function. A typical clinical assessment includes a perceptual evaluation, a patient history, a complete head and neck physical examination, acoustic or aerodynamic assessments, and a detailed laryngeal examination [4][16]. In general, the subjective nature of this evaluation can make assessment of the severity of a vocal disorder rather challenging. In particular, this complicates an evaluation of the efficacy of treatment.

Various researchers have suggested the need for standard battery of tests to evaluate a patient with spasmodic dysphonia [10][17][16]. A useful addition to such a battery would be a set of objective measures of vocal function derived from the speech waveform itself. Many researchers have examined the acoustic features of dysphonic phonation. Such features include shifts and deviations in fundamental frequency, signal aperiodicities, and voicing breaks [12][14][15]. These features, though, generally must be identified by hand, which introduces a measure of subjectivity into the process. Generally high inter-measurer correlations for these tasks have been reported [18], but the time and effort required for a human listener to identify such features is not insignificant. Automatic, objective methods for the identification of such features would reduce the time required for analysis while being highly repeatable.

In order to facilitate the assessment of patients with SD, this work proposes algorithmic methods of processing speech waveforms from patients with SD. Existing speech processing methodologies often rely on a voiced/unvoiced decision which is in turn based on the periodicity of a segment of the speech signal. Because of the noted prevalence of aperiodic segments of voiced speech in dysphonic phonation, the use of these standard techniques is not recommended. We propose an alternative approach that can identify both aperiodic segments and segments with stable fundamental frequency within voiced phonation. This approach allows the calculation of a wide variety of statistics which may assist in the discrimination of severity of dysphonia in a patient.

There is considerable disagreement regarding the use of sustained phonation versus continuous speech for an acoustic evaluation of the dysphonic voice. It is widely agreed that SD is significantly task dependent [11]. On the one hand, some researchers have suggested that sustained phonation produces more abnormal events and thus is more useful as a diagnostic tool [15]. Analysis of sustained phonation is also more straightforward because it provides an unambiguous context in which pathological features can be easily identified and relevant statistics can be easily calculated. On the other hand, continuous speech involves a wider variety of "speech tasks," which provides a greater cross section of the effects of a patient's dysphonia. Further, examining symptoms during continuous speech may provide a better indication of how severely a patient is effected during everyday use of the voice.

The ultimate goal of the present work is to develop discriminant statistics that correlate well with the perceived degree of dysphonia. We have chosen to compute these statistics from continuous speech recordings. During a six month study period, a database of recordings was collected from patients with SD reading a phonetically balanced passage. Each of these recordings was rated for degree of dysphonia by five trained listeners; these ratings serve as a baseline for evaluation of the discriminants presented here.

Man's First Boat

Long ago, men found that it was easier to travel on water than on land. They needed a clear path to travel on land, but on water a log of wood or any object that would float became a man's boat. It served to carry him across a stream or down a river.

Figure 1: The phonetically balanced passage used for vocal function assessment.

2 Methodology

2.1 Data

For this study, recordings were collected from thirty six patients with adductory spasmodic dysphonia over a six month period. Of the patients, thirty one were female, aged 27-85, and five were male, aged 35-67. The patients were asked to read a phonetically-balance passage, which is given in Figure 1. The recordings were made using a DAT recorder in a quiet room immediately prior to Botox injection. The database contains thirty seven recordings; one patient was given two injections during the six-month period and was thus recorded twice. The recordings were digitally transferred to a computer and resampled at 44,100 samples per second.

Each recording in the database was evaluated perceptually by five trained listeners: four speech language pathologists and one otolaryngologist. The recordings were presented in random order to each listener; the listeners were then asked to rate the extent of dysphonia for each recording on a one-hundred point scale, from "mild" to "severe." The resulting ratings were found to have a Kendall's coefficient of concordance [6] of 0.83, which indicates good agreement between the the listeners.

2.2 Preliminary investigation

The first stage of our investigation involved an examination of the recordings in our database. Since many of the commonly indicated pathological features of the dysphonic voice are related to the fundamental frequency (or, alternatively, the periodicity) of the speech signal, the use of a fundamental frequency tracking algorithm was a significant element of this investigation. The algorithm employed here is a modification of a time-domain autocorrelation approach presented by Boersma [2]. The details of this algorithm can be found in Appendix A at the end of this report. The algorithm computes the short-time autocorrelation of each of a set of overlapping frames of audio data, much like a spectrogram. This forms a surface with dimensions of time and lag. By locating peaks on the resulting surface, one can track the fundamental period (and thus, the fundamental frequency) of the signal over time. Additionally, the autocorrelation value at such a peak provides an indication of the extent of periodicity of that frame.

During the preliminary investigation, the recordings were examined by listening to them; through plots of the time-domain signal, the signal's estimated fundamental frequency, and the frame-correlation value; and through images of the

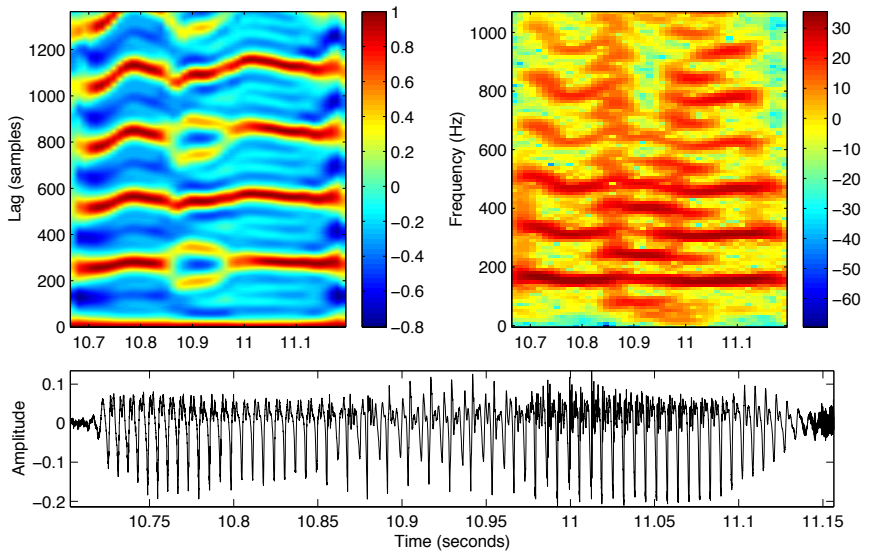

Figure 2: The autocorrelation surface (upper left), spectrogram (upper right), and time-domain waveform (bottom) for a signal that exhibits period doubling.

time-frequency and time-lag surfaces (via the spectrogram and the short-time autocorrelation, respectively). Some of the most relevant observations are recorded here.

2.2.1 Subharmonic events and period doubling

Previous investigations using this database have suggested that the appearance of subharmonics due to period doubling is a common pathological feature of SD. Our observations agree with this assessment. Figure 2 shows the time-domain signal, the autocorrelation surface, and the spectrogram for a speech signal that undergoes period doubling. Note that in this Figure the period doubling arises from the change in shape of alternate cycles of the signal. In general, these subharmonic events appear to be more common in mild to moderate cases of SD. For more severe forms, voicing is generally not stable enough for such subharmonic events to become evident.

2.2.2 Croak

Another common feature in our database is "croak" or "vocal fry." Croak appears as an abnormally low fundamental frequency during voiced segments of the speech signal. In some patients, the croak is a temporary change in voicing which results in a drop in fundamental frequency. In others, croak is evident throughout the recording. Figure 3 shows an example of a signal that exhibits croak. The segment shown begins with croaked phonation and then transitions to normal phonation. Note that this signal also involves a doubling of the signal's period. In Figure 3, however, the time-domain waveform shows a distinctive "damped oscillator" char-

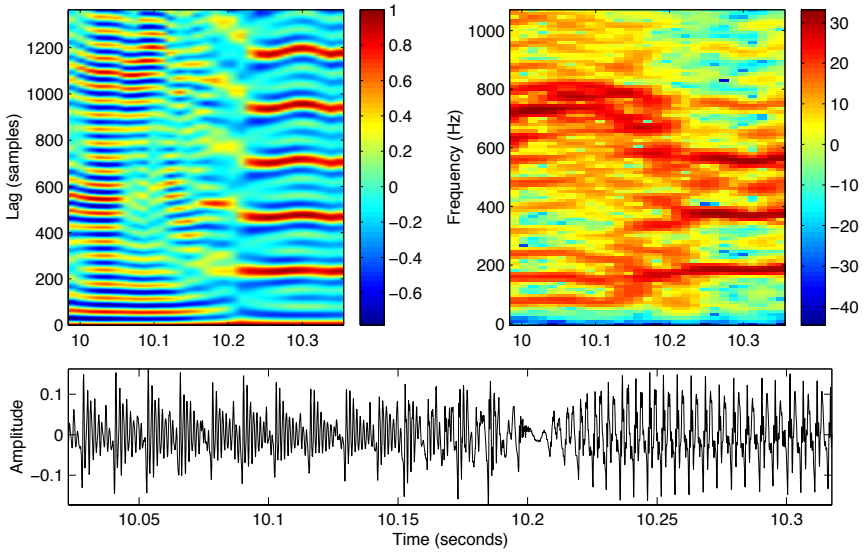

Figure 3: The autocorrelation surface (upper left), spectrogram (upper right), and time-domain waveform (bottom) for a signal that exhibits vocal croak.

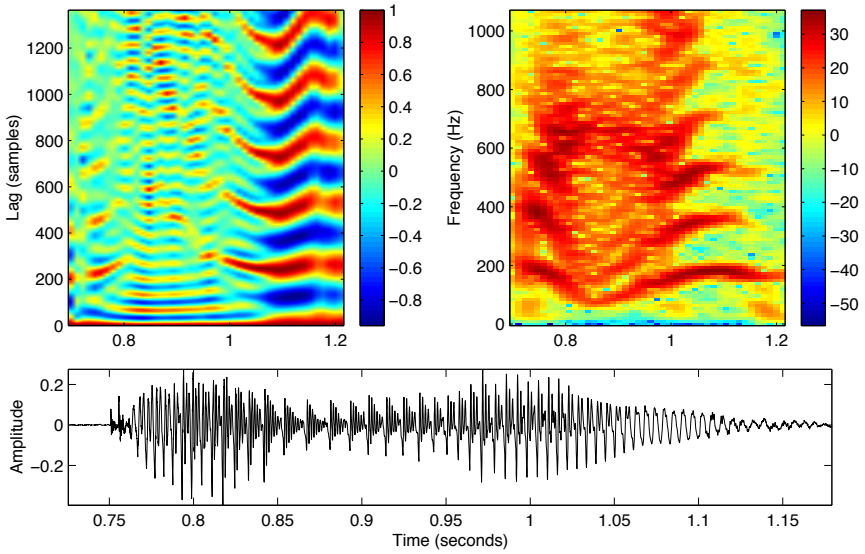

Figure 4: The autocorrelation surface (upper left), spectrogram (upper right), and time-domain waveform (bottom) for a signal that exhibits vocal croak due to a frequency shift.

acteristic, whereas in Figure 2 the period doubling results from a slight adjustment of the normal waveform shape. The autocorrelation surface in Figure 3 also shows a characteristic high frequency oscillation that is not evident in Figure 2. These differences suggest a different mechanism behind the two features.

In particular, croak is not restricted to period doubling. For instance, the signal shown in Figure 4 exhibits croak due to a rapid downward shift in fundamental frequency. This shift is not readily observable on the autocorrelation surface, but it can be visually inferred from the spectrogram. In the time-domain signal, individual cycles of the signal are visually apparent; however, frequency modulation changes the signal's period from cycle to cycle and prevents the identification of the signal's fundamental frequency using our autocorrelation method. Such "pseudo-periodicity" is a common feature of croaked phonation. The presence of a relatively strong first formant allows small changes in fundamental frequency to decorrelate successive cycles of the signal. Such small changes in fundamental frequency may arise from frequency shifts, vocal tremor, or simple prosody.

2.2.3 Voice breaks, aperiodicities, and frequency shifts

As suggested by other researchers, voicing breaks, voice tremor and frequency shifts, and aperiodic segments are also present in a significant number of recordings. Since our database consists of continuous-speech recordings, these features cannot be easily identified without contextual information about the utterance. Without such knowledge, these pathological features can easily be confused with common features of continuous speech. In continuous speech, for instance, voicing breaks typically result in the abnormal shortening of a voiced phoneme. Thus, a pathological voicing break is not significantly different from breaks that might occur between phonemes or words in normal speech. Similarly, voice tremor and frequency shifts can be difficult to distinguish from the normal frequency contours of speech prosody. Frequency shifts like the one in Figure 4 are often difficult to identify because the frequency modulation complicates the identification of the fundamental frequency. Finally, speech is naturally aperiodic at the boundaries between adjacent phonemes or words.

2.2.4 Voicing stability

One feature of the signals that was found to vary considerably across the database was the voicing stability. The voicing stability is the extent to which the speech signal adopts a continuous and stable period during voiced phonation. Informal listening seemed to indicate that voicing stability correlates well with the perceived severity of SD.

2.2.5 Clipping artifacts

One final important feature of the recordings in our database is not directly related to the patients' dysphonia. Many of the recordings in the database exhibit clipping artifacts. Often these artifacts are noise resulting from hard consonant

attacks or breath noise. In these cases, we would like to neglect frames that are corrupted by such clipping artifacts. In a few cases, though, the microphone level was not properly set, and significant portions of the voiced speech signal are clipped as a result. In these cases, the recording is still very understandable and clipped frames should not necessarily be neglected.

2.3 Processing of dysphonic speech

As indicated in the previous section, many of the pathological features of dysphonic speech are not easy to identify in the context of continuous speech. Further, dysphonic speech tends to violate some of the standard assumptions made in speech processing. Here, we propose an alternative approach for dysphonic speech that addresses these concerns. The algorithm is described below, and a more detailed presentation is given in Appendix B.

The literature suggests that the primary characteristics of dysphonic speech are variations in voiced speech. Thus, we propose that the first step in processing dysphonic speech, as with most other speech, is to perform voicing detection. However, as we have noted, we cannot use periodicity as our measure of voicing since we expect that some voiced phonation may be aperiodic. Instead, we adopt a voicing detection method that begins by identifying frames that are *not* voiced. The algorithm makes positive detections of silent, fricative, and clipped frames using simple rules based on easily-computed frame statistics. A silent frame is detected when a frame has an RMS value less than some fraction of the maximum signal value. Fricative frames are detected when a frame has a spectral centroid greater than some frequency. Clipped frame detection is based on the ratio of the frame's RMS value to the maximum value. To prevent the algorithm from improperly discarding voiced frames that are clipped, we only discard clipped frames if their peak autocorrelation falls below a threshold. Any frame that does not fall into one of these groups is considered to be voiced. We also remove a few frames at the edge of voiced segments to reduce the influence of aperiodicities due to starting and ending transients.

As in the preliminary investigation, a fundamental frequency tracking algorithm is an important part of our discrimination algorithm. The calculation of the short-time autocorrelation function is described in Appendix A. Once the short-time autocorrelation function has been calculated for each of a set of overlapping audio frames, a set of potential autocorrelation peaks is identified for each frame. These peaks are restricted to a particular range of the lag τ, which corresponds to an acceptable range of fundamental frequencies. Since a purely periodic signal will exhibit equal amplitude peaks for values of τ that are integer multiples of the fundamental period, a discounting factor is included to promote the selection of the lowest reasonable value of τ. The result should correspond to the fundamental period of the signal. Viterbi's algorithm is then used to search through the possible peaks for a "best path." Transition costs are determined as a function of the value of the short-time autocorrelation at the prospective peaks and the frequency difference between neighboring frames. The result is used as an estimate of the local

fundamental frequency for each frame of the speech signal.

The fundamental frequency estimate allows the identification of regions of stable voicing. In general, we consider that stable regions are separated by significant jumps in fundamental frequency. If these regions are too small, we neglect them as spurious. Formally, we say that a frame is *stably voiced* if it has not been classified as silent, fricative, or clipped *and* it belongs to one of a set of five or more consecutive frames not separated by a significant change in fundamental frequency. After this classification, any frame that has not yet been classified (as either unvoiced or as stably voiced) can implicitly be identified as voiced but either aperiodic or unstable. We will see that this "frame classification" approach allows the generation of statistics that correlate well with perceptual judgements of extent of dysphonia.

2.4 Deriving discriminant features

The "frame classifier" algorithm described in the previous section and in Appendix B provides a framework for computing a wide variety of discriminant features over a data set. Some such features are quite simple to compute; others can be significantly more complex. In general, we consider aggregate statistics for each recording; however, one could easily extend this method to the identification of specific pathological features in the speech signal. Here we describe some of the more successful discriminant features that we have investigated. In following sections, we evaluate these discriminant features. We also discuss methods of combining these features to produce a discriminant that is more highly correlated with the perceptual ratings for our database.

- **Unstable-to-voiced ratio.** One of the goals behind the frame classification algorithm is the identification of aperiodic segments within a speech signal. The unstable-to-voiced ratio provides an aggregate indication of how much voiced speech is either aperiodic or unstable due to frequency shifts or voice breaks.

- **Stable voicing fraction.** The stable voicing fraction is the ratio of the number of stably voiced frames to the total number of frames in the recording. This ratio provides an indication of the overall extent to which a passage is "stably voiced," which the preliminary investigation suggested correlates well with extent of dysphonia.

- **Mean voiced harmonic-to-noise ratio.** The mean voiced harmonic-to-noise ratio provides an indication of the overall periodicity of the voiced portions of the signal. Speech with aperiodicities and unstable fundamental frequencies have a lower voiced HNR.

- **Mean stable harmonic-to-noise ratio.** This ratio indicates the overall periodicity of stably voiced portions of a signal.

- **Signal length.** A statistically significant difference in mean time taken to speak a passage has been noted between dysphonic and normal speakers, and

between dysphonic patients before and after treatment with Botox injection [12]. This feature, however, will generally also depend on speech rate, and so it may not be a generally useful discriminant.

- **Mean stable length.** The "mean stable length" is the average number of frames in a "stably voiced" portion of the signal. This feature should be affected by voicing breaks and frequency shifts, and thus it may serve as an indication of the severity of dysphonia.

- **Mean frequency deviation.** The mean frequency deviation is the average magnitude change in fundamental frequency from frame to frame within a stably voiced segment. This feature should provide an indication of fluctuations in the fundamental frequency such as those produced by vocal tremor.

2.5 Discriminant evaluation

In order to evaluate the effectiveness of these SD discriminant features, we compare them to the ratings provided by five trained listeners. Each of the listener was asked to provided a score for "degree of dysphonia" on a one-hundred point scale. The perceptual nature of these ratings prevents us from comparing these scores directly among the listeners or between our algorithm and the the listener's scores. Instead, we consider that the only reliable information provided by these ratings is the *rank ordering* of the speech tokens.

In order to compare our discriminant features to the perceptual judgements of of these listeners, we rank-order the tokens based on the discriminant, and then compute the Spearman rank correlation [6], r_s, between these ranks and the ranks for each listener. An overall score for a particular discriminant is calculated as the mean value of the rank correlations with each listener. We have noted that Kendall's coefficient of concordance for the five listeners is equal to 0.83. This coefficient is approximately equal to the average *pairwise* rank correlation over all pairs of judges [6]. Thus, a discriminant with mean rank correlation of approximately 0.83 would agree with the judges as well as the judges agree with themselves.

3 Results

We have organized the evaluation of results from our algorithmic methods into three "experiments." The first experiment evaluates the use of each feature described in Section 2.4 as a discriminant for determining the extent of dysphonia. The second experiment examines ways of combining multiple features to produce a combined discriminant. The third experiment examines how well the combined-features method operates when "trained" and "tested" on separate data sets.

3.1 Experiment #1: Feature Evaluation

Table 1 shows the mean rank correlation for each of the features provided in Section 2.4. For $N = 37$, we reject the null hypothesis of statistical independence at

#	Feature	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Mean
1	U-to-V Rat	0.7037	0.5967	0.5510	0.6265	0.7611	0.6478
2	Stable VF	-0.7988	-0.7859	-0.7350	-0.6839	-0.8156	-0.7638
3	Voiced HNR	-0.6687	-0.5638	-0.5165	-0.5851	-0.6709	-0.6010
4	Stable HNR	-0.5342	-0.4446	-0.3783	-0.5951	-0.5557	-0.5016
5	Signal Leng	0.4364	0.5268	0.5912	0.3935	0.4655	0.4827
6	Stable Leng	-0.6105	-0.4889	-0.4230	-0.4419	-0.6288	-0.5186
7	Frq Dev	0.6430	0.5032	0.5056	0.5974	0.6869	0.5872

Table 1: Spearman rank correlations between each of seven features and the five judges.

$p \leq 0.01$ for rank correlation values of $|r_s| \geq 0.42$. Nearly all of the features presented here show statistically significant rank correlations with each of the judges.

The feature with the largest mean rank correlation is the "stable voicing fraction" with a mean rank correlation of -0.7638. Since only the rank order of the discriminant contributes to the rank correlation, we can easily negate or invert the discriminant to achieve a positive correlation. Thus, this correlation indicates high correlation with the perceived extent of dysphonia. It does, however, fall below the mean inter-judge rank correlation of 0.83. The "unvoiced-to-voiced ratio" and "voiced HNR" also show good correlation with a mean rank correlation of 0.6478 and -0.6010, respectively.

3.2 Experiment #2: Combining features into a single discriminant

Each of these features captures a somewhat different aspect of the the underlying signal. Thus, one might expect that combining the information from several features into a single discriminant might produce a discriminant with improved performance. Here, we examine the usefulness of producing a new discriminant, d, as a linear combination of N calculated features, f_i, as

$$d = \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_N f_N \tag{1}$$

using weights α_i. For simplicity, we normalize the features to have zero mean and unit variance.

The primary challenge with this method is identifying weights that yield a discriminant with improved correlation. This requires a training procedure to identify appropriate weights. We have examined two such procedures. The first involves a randomized search of the parameter space. In this method, a number of random weights are generated, and the mean rank correlation between the resulting discriminants and the listener's ratings are computed. The sets of weights with maximum mean rank correlation are then perturbed randomly in a manner similar to simulated annealing. This procedure, while somewhat slow, can produce parameter settings that produce discriminants with rather high mean rank correlation. Because of its speed, however, using this method to identify an appropriate set of features (that is, using it for feature selection) can be quite laborious.

Feature set	MMSE r_s	Refined r_s
[2 5]	0.7909	0.8004
[2 6]	0.7743	0.7796
[2 3]	0.7629	0.7677
[2 5 6]	0.7954	0.8037
[2 3 5]	0.7920	0.8032
[1 2 5]	0.7906	0.8028
[2 5 6 7]	0.8032	0.8156
[1 2 3 5]	0.7983	0.8051
[2 3 5 6]	0.7959	0.8084

Table 2: Three best feature sets for two, three, and four features. The resulting least-squares and refined rank correlations are listed.

A second procedure produces slightly lower mean rank correlations, but is significantly faster and highly repeatable. This procedure involves projecting a "target" vector with one dimension per recording in the database onto the subspace spanned by the N discriminants. That is, if $\mathbf{b}$ is the target vector and A is a matrix with the set of N f_i's as columns, then we are seeking a least-squares solution to the overdetermined system defined by

$$A\mathbf{x} = \mathbf{b}. \tag{2}$$

Identifying an appropriate target vector is not trivial because our "target" is truly ordinal in nature. However, we have found that good results are achieved by using a vector derived from the ranks given by the trained listeners. To form this target vector, we take the mean of the listener's ranking vectors, rank order the result, and subtract the mean. With such a target vector, a least squares solution may be readily obtained. The speed of this procedure suggests that we might use the least-squares technique for the selection of an appropriate feature set, and then refine the results using the randomized optimization method.

For our second experiment, we have used the least-squares technique to identify the three best feature sets with two, three, and four features. These feature sets were then refined using the random optimization method described above. Table 2 presents the results. The addition of a second feature increases the mean rank correlation from 0.76 to 0.80. Additional features beyond two only increase the rank correlation slightly beyond this to a maximum of 0.8156 with four features. These scores suggest that we are at a knee in the performance curve and that further gains may require correspondingly more features. The rank correlations we obtain here are better than the best single-feature rank correlation, but is still below the mean inter-judge rank correlation of 0.83.

The vector-space interpretation of the search for a combined discriminant suggests a potential problem with this approach. In general, as we introduce more features, we are able to produce discriminants that better match the the perceptual rankings. However, these features do not necessarily need to correlate with spasmodic dysphonia. Any linearly independent (but potentially random!) feature has

the potential to improve correlation because of the corresponding increase in the dimension of the subspace spanned by the features. In the limit, if we have one (linearly independent) feature for each recording in our database, we can match our target vector perfectly, even if our features are completely random. In such a case, we will have strongly overtrained our system, and its performance on other data is likely to be significantly degraded. Because of this, we have restrict our attention to small feature sets for this experiment. In the next section, we examine how well this method operates when trained and tested on different sets of data.

3.3 Experiment #3: Training and testing on different data sets

In the previous experiments, we identified multiple-feature discriminants by "training" a set of weights for linear combination of the features. As was indicated, we must be careful that we are not overtraining our discriminants on the present data, thus inflating the results obtained on this data set at the expense of results on a larger data set. To examine the wider applicability of our techniques, here we will examine the results when we train on one subset of the database and test on another (disjoint) subset.

The procedure for this experiment is as follows. We first partition the database into multiple pairs of training and testing sets. The training and testing sets that form one pair are disjoint and each contain 18 recordings from the original database. To assure a reasonable distribution between the training and testing sets, each partition is chosen so that the mean rank of the two sets is approximately equal. For each partition, we "train" four discriminants, one each with one, two, three, and four features. The discriminants are trained by selecting the "best" feature set of all sets with the desired number of features. The "best" feature set has the highest mean rank correlation as identified using the least-squares technique described in the previous section. The resulting feature weights are then used to compute discriminants for the testing set, and the mean rank correlation is computed for this set.

Table 3 shows the results of this procedure for four partitions of the database. We can note from this data that there is significant variability between the rank correlation for the training set and the testing set. Despite this variability, though, most of the testing set correlations are still fairly high, especially in the first two partitions. For these examples, it is not clear that the use of additional features to improve the rank correlation of the training set actually improves the rank correlation for the testing set. In fact, the evidence suggests that additional features may actually degrade performance. Except for the first partition, the testing set performance is the best with the single-feature discriminant.

4 Discussion

The results of our three experiments have shown that we can produce a discriminant that correlates well with perceptually-rated degree of dysphonia. In particular, the first experiment indicates that the "stable voicing fraction" that we have em-

	Features	Train r_s	Test r_s
Partition #1	[2]	0.7415	0.7288
	[2 5]	0.8410	0.7672
	[2 5 7]	0.8463	0.7397
	[2 5 6 7]	0.8430	0.7393
Partition #2	[2]	0.6827	0.8303
	[2 5]	0.7429	0.8208
	[2 3 5]	0.7504	0.8208
	[1 2 3 5]	0.7504	0.8054
Partition #3	[2]	0.7107	0.7881
	[2 5]	0.7658	0.7857
	[1 2 7]	0.8247	0.5828
	[1 2 4 7]	0.8304	0.5619
Partition #4	[2]	0.7750	0.6869
	[2 5]	0.8021	0.6782
	[2 3 5]	0.7990	0.6446
	[2 3 5 6]	0.8033	0.6202

Table 3: Rank correlations for four training/testing partitions of the database.

ployed here is a strong indicator of severity of dysphonia. Some of the other features, such as the "unstable-to-voiced ratio" and the "mean harmonic-to-noise ratio" also serve as good indicators of severity. Still, none of our single feature discriminants agree with the judges as well as the judges agree with one another.

The second experiment indicates that we can improve discriminant correlation somewhat by combining multiple features into a single discriminant. In doing so, however, we must be mindful of the consequences for the discriminant's performance on a larger data set. The data for the second experiment suggests that the best tradeoff between high correlation and low complexity is a simple two-feature discriminant.

The third experiment shows that a multi-feature discriminant trained on one half of our database does not necessarily produce improved results on the other half. This throws doubt on whether the multi-feature discriminants generated in the second experiment will actually generalize well to a larger data set. The feature-combination process is inherently statistical, and our database may not be large enough to produce results that can generalize to a larger population of patients.

We can obtain another indication of the general applicability of a particular discriminant by testing it on recordings of speakers who do not have spasmodic dysphonia. To this end, the seven features described in 2.4 were computed for three "control" recordings of the author reading the "Man's First Boat" passage. The resulting rank (out of 40) for each recording is presented in Table 4. Recall that features 1, 5, and 7 have a positive correlation with SD extent, so the control recordings should have very low ranks; features 2, 3, 4, and 6 exhibit a negative correlation so we expect high rankings. Some of the features, such as the mean frequency deviation, do show the expected trend. For others, these "control" recordings are ranked

#	Feature	Control #1	Control #2	Control #3
1	U-to-V Rat	7	1	3
2	Stable VF	14	25	23
3	Voiced HNR	31	38	35
4	Stable HNR	21	30	27
5	Signal Leng	4	17	1
6	Stable Leng	28	35	33
7	Frq Dev	1	4	2

Table 4: Rankings (out of 40) for three "control" recordings made by the first author.

somewhere near the middle of the database, which seems to suggest that the author is afflicted by a moderate form of spasmodic dysphonia. Particularly interesting is the fact that the stable voicing fraction, which achieved very good correlation with SD extent, does not rate the control recordings as "better" than most of the recordings in the database. These preliminary suggest that mild dysphonia may be characterized by more subtle features than the aggregate statistics we have computed here can capture. They further suggest the need for a more detailed study of these statistics for dysphonic and non-dysphonic patients.

While we certainly would not suggest that the discriminants identified in this study form the sole basis of an assessment of patients with SD, such scores could potentially be used as one component of the clinical assessment. Because these discriminants are obtained algorithmically, they are less subjective than similar scores provided by a human observer trying to identify aperiodic segments or frequency shifts. Further, we suggest that the framework presented here is generally useful for the processing of dysphonic speech. One could use it to explore the correlation between various features of the waveform and characteristics such as breathiness or strain-strangled quality. More advanced features that incorporate contextual information might also be developed. In particular, additions such as a text alignment system or a prosody model might prove useful for examining and identifying aspects of spasmodic speech, such as voicing breaks or vocal tremor.

5 Conclusion

In this study, we have presented a system for processing continuous-speech recordings of dysphonic speech. This system classifies frames of audio as stably voiced, unstable/aperiodic, silent, fricative, or clipped while also providing a fundamental frequency estimate and a periodicity rating for each frame. The information returned by the system allows the calculation of various statistics from continuous speech recordings. We have found that one such statistic in particular, the "stable voicing fraction," correlates very well with the degree of dysphonia as rated by five trained observers. Other features correlate also exhibit significant correlations with extent of dysphonia. Methods of combining features to produce a more highly correlated discriminant are examined, but the reliability of such discriminants on

larger data sets is uncertain.

Acknowledgements

I am grateful for the assistance of Dr. Norman Hogikyan for the collection of the data used in this work. I would also like to thank my research advisor, Dr. Gregory Wakefield, for his assistance, direction, and helpful suggestions.

A Appendix: Periodicity analysis

Determining the fundamental period (and thus, the fundamental frequency) of a signal is a common task for speech and audio processing, and a significant body of research has been devoted to this problem. Hess [8] and Hermes [7] provide a good overview of these so-called "pitch detection" algorithms. In some cases, we may also be interested in the extent to which a signal varies from perfect periodicity over a short time frame. Unfortunately, most fundamental frequency estimation algorithms do not also indicate the extent of periodicity. More recently, algorithms have been presented that do provide this information. One such algorithm is presented by Boersma [2]. In this work, we employ a modification of Beorsma's method for periodicity detection.

A.1 Short-term autocorrelation and Boersma's method

Most time-domain, autocorrelation-based fundamental frequency tracking algorithms operate on the basis of an analogy to the autocorrelation of random processes which are wide sense stationary and ergodic. For the WSS, ergodic random process $\mathbf{x}(t)$, we compute the normalized autocorrelation function $r_x(t)$ as,

$$r_x(\tau) = \frac{E\big[\mathbf{x}(0)\mathbf{x}(\tau)\big]}{E\big[\mathbf{x}^2(0)\big]} = \frac{\int \mathbf{x}(t)\mathbf{x}(t-\tau)dt}{\int \mathbf{x}^2(t)dt}, \tag{3}$$

where $E[\cdot]$ indicates the expectation of a random variable. The second equality in (3) arises from the assumption of ergodicity in correlation. $r_x(\tau)$ has a global maximum of $r_x(\tau) = 1$ at $\tau = 0$. If $\mathbf{x}(t)$ is a periodic process with fundamental period T_0, then $r_x(\tau)$ will also have global maximum at each $\tau = kT_0$, where $k \in \mathbb{Z}$. By locating maxima of $r_x(\tau)$, one can identify the fundamental period (and thus the fundamental frequency) of $\mathbf{x}(t)$.

For pseudo-periodic processes, the value of the autocorrelation function is also useful for determining "how periodic" the process is. Suppose that we let $\mathbf{y}(t) = \mathbf{x}(t)+\mathbf{v}(t)$, where $\mathbf{x}(t)$ is periodic with period T_0 and $\mathbf{v}(t)$ is a white noise process that is uncorrelated with $\mathbf{x}(t)$. The autocorrelation function of the sum of uncorrelated processes is the sum of the autocorrelations, so

$$r_y(\tau) = \frac{r_x(\tau) + N_0\delta(\tau)}{N_0 + r_x(0)}, \tag{4}$$

where N_0 is the power of $\mathbf{v}(t)$. From this, we can determine the *harmonic to noise ratio*, which is defined as the power of the "periodic part" of $\mathbf{y}(t)$ to the "non-periodic part," as

$$\text{HNR} = \frac{r_y(T_0)}{1 - r_y(T_0)} = \frac{r_x(0)}{N_0}. \tag{5}$$

Thus, the value of $r_y(\tau)$ at $\tau = T_0$ acts as an indicator of the periodicity $\mathbf{y}(t)$.

For real applications of a periodicity detector, we generally need to estimate the fundamental period of a non-stationary signal that may be nearly periodic over some finite support. To do this, we apply a symmetric, finite-support window $w(t)$ to the signal $x(t)$ before calculating the autocorrelation. This yields the normalized short-term autocorrelation function, $r'_x(\tau)$, which is defined as

$$r'_x(\tau) = \frac{\int x(t)w(t)x(t-\tau)w(t-\tau)dt}{\int x^2(t)w^2(t)dt}. \tag{6}$$

Here, if $x(t)$ is locally periodic with period T_0 over the time support of $w(t)$, then $r'_x(\tau)$ will have a relative maxima at $\tau = T_0$. Unfortunately, the window function also introduces a taper into $r'_x(\tau)$, which complicates the identification of the correct lag value for the signal's fundamental period [13].

Boersma suggests that this taper can be eliminated by dividing $r'_x(\tau)$ by the autocorrelation of $w(t)$. No justification is provided; however, if we view $r'_x(\tau)$ as an estimator of $r_x(\tau)$, we can see that the expected value of $r'_x(\tau)$ is

$$E\left[r'_x(\tau)\right] = \frac{\int w(t)w(t-\tau)dt}{\int w^2(t)dt} r_x(\tau). \tag{7}$$

Thus, $r'_x(\tau)$ is a biased estimator of $r_x(\tau)$, with a bias equal to the normalized autocorrelation of $w(t)$. Dividing by the the autocorrelation function of the window produces the unbiased estimator

$$\hat{r}'_x(\tau) = \frac{r'_x(\tau)}{r_w(\tau)}. \tag{8}$$

Note that the variance of this estimate can become significant for large values of τ [9]. The resulting estimate of the long-term autocorrelation function proves to be a useful method of detecting local periodicity in a signal, providing both an estimate of the period and the degree of periodicity.

A.2 Extensions to Boersma's method

Boersma's method, while quite useful, is not perfect. In particular, consider the use of (8) to calculate the harmonic to noise ratio. For this calculation to be meaningful, $\hat{r}'_x(\tau)$ must be no greater than one. It turns out, however, that (8) is not guaranteed to remain less than one. Boersma suggests that autocorrelation values that are greater than one should be "reflected through 1" before computing the harmonic to noise ratio. Again, no justification is provided for this procedure, and it's use is somewhat questionable. By modifying the calculation of our estimated

autocorrelation, however, we can guarantee that our short-term autocorrelation will never exceed one.

The basic idea is that we compute the normalized correlation between a windowed signal and its shifted counterpart. That is, we want our candidate autocorrelation to have the form

$$\tilde{r}'_{z_1,z_2}(\tau) = \frac{\langle z_1(t), z_2(t)\rangle}{\|z_1(t)\|\|z_2(t)\|}. \tag{9}$$

Then, by the Cauchy-Schwarz inequality, $\tilde{r}_{z_1,z_2}(\tau) \leq 1$. Note that (6) already has this form. However, we would like to meet another constraint as well. When the underlying signal, $x(t)$ is periodic with period T_0, we want $\tilde{r}_{z_1,z_2}(T_0) = 1$. This ensures that the harmonic to noise ratio calculation given in (5) is still valid. One way to satisfy these constraints is to use

$$z_1(t) = x(t)\sqrt{w(t)w(t-\tau)} \tag{10}$$
$$z_2(t) = x(t-\tau)\sqrt{w(t)w(t-\tau)}, \tag{11}$$

where w(t) is a symmetric window. If $x(t) = x(t - T_0)$, then $z_1(t) = z_2(t)$ and $\tilde{r}'_{z_1,z_2}(\tau)$ reduces to 1. This yields a first alternate form for the local estimate to the autocorrelation, $\tilde{r}'_x(\tau)$, which is given by

$$\tilde{r}'_x(\tau) = \frac{\int x(t)w(t)x(t-\tau)w(t-\tau)dt}{\sqrt{\int x^2(t)w(t)w(t-\tau)dt \int w(t)x^2(t-\tau)w(t-\tau)dt}}. \tag{12}$$

Both (12) and (8) are nonideal in one important respect. In both of these calculations, the estimated autocorrelation values for increasingly large values of τ are computed with progressively less data support. This occurs because the non-zero portion of the effective window $w(t)w(t-\tau)$ shrinks as τ increases. Because of this, Boersma suggests that the maximum useful value of τ is one-third of the window length. In particular, $\hat{r}'_x(\tau)$ and $\hat{r}'_x(\tau)$ are undefined for τ greater than the window length. We can alleviate these problems by defining the short-term autocorrelation function such that it uses a constant window for all τ. This suggests a second alternative form of the short-term autocorrelation function, given by

$$\bar{r}'_x(\tau) = \frac{\int x(t+\tau/2)w(t)x(t-\tau/2)dt}{\sqrt{\int x^2(t+\tau/2)w(t)dt}\sqrt{\int x^2(t-\tau/2)w(t)dt}}. \tag{13}$$

There are several important aspects of this equation. First, we are windowing the product of shifted versions of $x(t)$ rather than simply multiplying shifted versions of $x(t)w(t)$. This provides a single window for all τ, as desired. Second, we have introduced a different time shift to both "copies" of $x(t)$ and the window $w(t)$. This is necessary because, to prevent the introduction of a lag-dependent delay, the effective window of $x(t)x(t-\tau)$ must be centered at $\tau/2$. In (12) and (8), this was achieved by using $w(t)w(t-\tau)$ as the effective window. Here, we achieve this goal by shifting $x(t)$ in opposing directions while leaving $w(t)$ "stationary."

We suggest that (13) is an ideal theoretical definition for the short-term autocorrelation for several reasons. First, as with $\tilde{r}'_x(\tau)$, $\bar{r}'_x(T_0) = 1$ when $x(t)$ is periodic with period T_0. Second, by construction we know that $\bar{r}'_x(\tau) \leq 1$, which allows (5) to always be meaningful. Third, the computation uses a uniform support size for all τ. Finally, $\bar{r}'_x(\tau)$ is produces a meaningful result for *all* τ, regardless of the window length. Unfortunately, as we will see in the next section, $\bar{r}'_x(\tau)$ requires significantly more computation than either $\tilde{r}'_x(\tau)$ or $\hat{r}'_x(\tau)$.

A.3 Implementation Issues

Though developed in continuous time, the various short-term autocorrelation functions presented in the previous sections can readily be implemented in discrete time. The primary concern in doing so is the loss of resolution in τ when identifying peaks on the short-term autocorrelation function. Assuming that all signals are bandlimited, we can ideally reconstruct the continuous-time short-term autocorrelation function from the discrete-time short-term autocorrelation function through sinc interpolation. In particular, Boersma suggests using sinc interpolation to refine estimates of maxima on the autocorrelation function. In practice, we have found that using parabolic refinement of local maxima actually produces better accuracy with significantly less computation.

Another potential difficulty that arises from sampling occurs in the definition of $\bar{r}'_x(\tau)$ (13). Here, we are shifting $x(t)$ by the value $\tau/2$. To implement this in discrete time, though, we must shift by an integer number of samples. By either upsampling our signals by a factor of two or by assuring that they are bandlimited to $f_s/4$, we can replace $\tau/2$ by an integer shift. Then, we can perform the computation without aliasing. This is effectively the same operation required when computing the discrete-time pseudo-Wigner distribution [3].

$\hat{r}'_x(\tau)$ can be computed efficiently by using the fast Fourier transform to compute the necessary correlations. Since the autocorrelation function of the window can be computed off-line, the algorithm requires only two FFTs for each short-term autocorrelation function that we wish to calculate. To prevent artifacts from the circularity of DFT-based convolution, it is necessary for the DFT length to be greater than 1.5 times the window length [2]. The computation of $\tilde{r}'_x(\tau)$ can be computed in a similarly efficient way. In this case, however, we need four FFTs per frame – two each for the correlations in the numerator and the denominator of (12). (The results of the two correlations in the denominator are time-reversed versions of one another, so we only need to compute one of them.) The computation of both forms of the short-term autocorrelation function has a complexity of $O(N \log_2 N)$ operations per frame, where N is the DFT length.

Our "ideal" short-term autocorrelation function, $\bar{r}'_x(\tau)$, is not so efficient to compute. The numerator of (13), with its two distinct time shifts, cannot be expressed as a simple correlation. This further means that we cannot perform this calculation with the fast Fourier transform. The resulting computational complexity is $O(MN)$ per frame, where N is the window length and M is the number of values of τ that we evaluate. For the present application, we are computing autocorrelations for one

hundred frames of data per second over 20 to 40 seconds; thus, the computational requirements for $\bar{r}'_x(\tau)$ become rather significant.

Because of the computational complexity required for $\bar{r}'_x(\tau)$, we employ the short-time autocorrelation function given by $\tilde{r}'_x(\tau)$ for periodicity detection in this study. We compute $\tilde{r}'_x(\tau)$ as

$$\tilde{r}'_x(\tau) = \frac{\mathrm{DFT}^{-1}\Big\{ \big|\mathrm{DFT}\{x[n]w[n]\}\big|^2 \Big\}}{\sqrt{\mathrm{DFT}^{-1}\Big\{\mathrm{DFT}\{x^2[n]w[n]\}W^*[k]\Big\}}\sqrt{\mathrm{DFT}^{-1}\Big\{\mathrm{DFT}\{x^2[n]w[n]\}^*W[k]\Big\}}}, \tag{14}$$

where $x[n]$ is one frame of the input signal, $w[n]$ is the window function, and $W[k]$ is the DFT of the window function. As previously noted, the terms in the denominator are time-reversed versions of one another, so the entire computation requires only four DFTs per frame.

B Appendix: Frame classification algorithm

Here, we describe the frame classification algorithm that forms the basis of our discriminant calculation method.

1. **Define a frame segmentation.** Following [2], we use a frame size which is equal to three times the period of our minimum expected fundamental frequency. We use a minimum fundamental of 50 Hz, which yields a frame size of 60 ms. Our frame step size is 10 ms.

2. **Filter the input signal.** As a preprocessing step, we lowpass filter the input signal to 900 Hz. This reduces the susceptibility of the periodicity detection to frequency modulation.

3. **For each frame:**

   (a) **Calculate the short-time autocorrelation.** Using (14), we compute the short-time autocorrelation function of one frame of the lowpass filtered signal. We use an FFT length equal to the smallest power of two greater than the window size.

   (b) **Locate peaks of the autocorrelation function.** We identify the six most likely peaks within the range of lags corresponding to a frequency range of 50 Hz to 400 Hz. Following [2], the "most likely" peaks are determined subtracting a 0.02 per octave discounting factor from the autocorrelation function's value. The location and height of each peak is refined using parabolic interpolation.

   (c) **Calculate frame statistics.** The RMS value, spectral centroid, and the maximum (windowed) value are computed for the frame using the original (unfiltered) signal.

4. **Identify best frequency path.** Using the Viterbi algorithm, we identify the lowest-cost path through the potential peaks of the autocorrelation surface. The cost of a transition is equal to 0.2 times the number of octaves jumped less the value of the autocorrelation function at the next frame.

5. **Identify unvoiced frames.** Unvoiced frames are identified using the following rules:

   - A frame is *fricative* if its spectral centroid is greater than 400 Hz.
   - A frame is *silent* if its RMS value is less than 0.01 times the maximum signal value.
   - A frame is *clipped* if the ratio the the frame's maximum value to its mean squared value is less than 10 *and* the frame's peak autocorrelation value is less than 0.8.

   Any frame that is classified as one of the above is considered to be unvoiced. We perform three binary erosions and one binary closure on the remaining set of voiced frames to remove frames that transition from voiced to unvoiced.

6. **Identify stably voiced frames.** In this step, we collect consecutive frames into groups of stably voiced frames. These groups are separated by jumps in fundamental frequency in which the lower of the two neighboring frequencies is less than 0.65 times the higher. Any such group that contains fewer than 5 frames is considered to be unstable.

7. **Calculate features.** Having classified each frame of the audio signal, we can now calculate features as candidate discriminants for dysphonic severity. There is a wide array of potential discriminants; some of these are discussed in Section 2.4.

References

[1] A. Blitzer and M. F. Brin. The dystonic larynx. *Journal of Voice*, 6(4):294–297, 1992.

[2] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 17, pages 97–110, 1993.

[3] T. A. C. M. Claasen and W. F. G. Mecklenbrauker. The wigner distribution – a tool for time-frequency nalaysis, part ii: Discrete-time signals. *Philips Journal of Research*, 35(4/5):277–300, 1980.

[4] R. H. Colton and J. K. Casper. *Understanding Voice Problems.* Williams and Wilkins, Baltimore: Maryland, 1990.

[5] C. N. Ford, D. M. Bless, and N. Y. Patel. Botulinum toxin treatment of spasmodic dysphonia: Techniques, indications, efficacy. *Journal of Voice*, 6(4):370–376, 1992.

[6] W. L. Hayes. *Statistics for the Social Sciences*. Holt, Rinehart and Winston, Inc., Upper Saddle River, New Jersey, 1973.

[7] D. J. Hermes. *Pitch analysis*, chapter 1, pages 3–25. John Wiley & Sons Ltd, 1993.

[8] W. J. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.

[9] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing*. Prentice Hall PTR, New York, 1973.

[10] K. Izdebski. Symptomatology of adductor spasmodic dysphonia: A physiologic model. *Journal of Voice*, 6(4):306–319, 1992.

[11] C. L. Ludlow and N. P. Connor. Dynamic aspects of phonatory control in spasmodic dysphonia. *Journal of Speech and Hearing Research*, 30:197–206, June 1987.

[12] C. L. Ludlow, R. F. Naunton, S. E. Sedory, G. M. Schulz, and M. Hallett. Effects of botulinum toxin injections on speech in adductor spasmodic dysphonia. *Neurology*, 38:1220–1225, 1988.

[13] L. R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25:24–33, 1977.

[14] C. M. Sapienza, T. Murray, and W. S. Brown. Variations in adductor spasmodic dysphonia: Acoustic evidence. *Journal of Voice*, 12(2):214–222, 1998.

[15] C. M. Sapienza, S. Walton, and T. Murry. Acoustic variations in adductor spasmodic dysphonoia as a function of speech task. *Journal of Speech, Language, and Hearing Research*, 42(1):127–140, Feb 1999.

[16] C. F. Stewart, E. L. Allen, P. Tureen, B. E. Diamond, A. Blitzer, and M. F. Brin. Adductor spasmodic dysphonia: Standard evaluation of symptoms and severity. *Journal of Voice*, 11(1):95–103, 1997.

[17] G. E. Woodson, P. Zwirner, T. Murry, and M. R. Swenson. Functional assessment of patients with spasmodic dysphonia. *Journal of Voice*, 6(4):338–343, 1992.

[18] P. Zwirner, T. Murry, and G. E. Woodson. Perceptual-acoustic relationships in spasmodic dysphonia. *Journal of Voice*, 7(2):165–171, 1993.