INDIVIDUAL VARIATIONS IN GLOTTAL CHARACTERISTICS OF FEMALE SPEAKERS

Helen M. Hanson

Harvard University, Division of Applied Sciences
Cambridge, MA 02138, USA

ABSTRACT

We address the measurement of glottal characteristics of female speakers and how these characteristics contribute to voice quality or individuality. We have developed acoustic measurements of the voicing source that are made directly on the speech waveform or spectrum. These measures are based on theoretical models of speech production. Based on these measurements it is possible to make some inferences about the glottal configuration during phonation and the nature of the glottal pulse. Previous work has relied mainly on physiological methods such as inverse filtering of vocal tract airflow or observations of vocal fold vibration via endoscopy or fiberscopy. Our measures are non-invasive and can be easily extracted automatically. By comparing these acoustic measures to physiological and perceptual data, we show that they are valid.

1. INTRODUCTION

Glottal source characteristics have been shown to contribute to voice quality and capture information about speaker individuality [1], suggesting that the ability to quantify and measure these characteristics is of importance to speech processing. Speaker recognition or verification, speech synthesis, and even speech recognition can be improved with better knowledge of voice source characteristics. There are also medical applications, such as diagnosis and classification of voice or speech disorders, and biofeedback systems used in speech or voice therapy. In addition, there are forensic applications.

In our work we attempt to formulate quantitative measures of individual differences that can occur in the spectrum of the glottal pulses for normal adult female speakers. Previously, glottal characteristics have been studied using inverse filtering and physiological measurements [2], [3]. Methods such as these are not practical for most of the applications mentioned above. However, theoretical models of speech production can be used to predict how differences in glottal behavior show up in the acoustic sound pressure [4], [5]. We use measurements on the speech waveform and spectrum as the basis for classification of individual glottal differences. These measures are noninvasive and should be simple and efficient to extract automatically by computer.

We begin the paper by first reviewing our measurements and previous results based on 13 speakers [5]. Then we present new data based on increasing our subject pool to 22 speakers. We also report the results of listening tests and physiological data, and relate these data to our acoustic measurements.

2. ACOUSTIC MEASUREMENTS

In [5] we developed the following acoustic measurements based on theoretical models of speech production:

First-formant bandwidth. When the glottis is not closed during phonation, the resulting glottal resistance can contribute to energy losses in the frequency range of F1. F1 bandwidth can indicate the size of the glottal opening [6]. We estimate bandwidth by bandpass filtering the speech waveform around F1 and measuring the decay of the resulting waveform during the initial part of the glottal cycle.

H1*-H2* (difference in spectral amplitude of the first and second harmonics). As the open quotient (OQ) of a glottal waveform increases, the source spectrum undergoes a change primarily at the low frequencies, and thus we use H1-H2 as a measure of OQ. Corrections are made for the amounts by which H1 and H2 are "boosted" by F1, yielding H1*-H2*.

H1*-A1 (where A1 is the amplitude of the strongest harmonic in F1). The prominence of the F1 peak may reflect the formant bandwidth and can also be reduced by a large spectral tilt. We use H1*-A1 (with H1 corrected for the influence of F1) as an indication of presence and size of a glottal opening.

H1*-A3* (where A3 is the amplitude of the strongest harmonic in F3). The glottal spectrum at middle and high frequencies (spectral tilt) is mostly influenced by how abruptly the flow is cut off when the glottis closes [4]. This abruptness is affected by the presence of a glottal opening during the closed phase, by whether or not the folds close simultaneously along their length, and by how rapidly the folds close. We use H1-A3 as a measure of spectral tilt, where H1 is corrected as above, and A3 is corrected for the effect of F1 and F2 on the spectrum amplitude of F3, yielding H1*-A3*. The value of H1*-A3* enables us to estimate how the folds are closing.

Noise ratings. An acoustic consequence of a glottal opening during phonation is the generation of turbulence noise in the vicinity of the glottis. We use the technique of

The author is also with and can be reached at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139

This work is supported in part by NIH grants DC00075 and DC002661

Klatt and Klatt [4] to rate the amount of noise in the F3 waveform (obtained by bandpass filtering) and in the corresponding spectrum. Estimates are made on a scale from 1 to 4, where 1 means there is no evidence of noise interference and 4 means that there is little evidence of periodicity. The amount of noise can then used to estimate the size of a glottal opening. We have found that the waveform-based rating has a high correlation to the spectrum-based rating.

Stevens and Hanson [5] presented the results of applying these measurements to steady-state vowels excised from the running speech of a group of 13 female speakers. The measurements showed substantial variations across speakers. We were able to make some inferences about glottal configurations of the subjects during phonation. The speakers fell into two groups. One group (Group 1) had nine members with low spectral tilt, evidence that they maintain closure at the vocal processes. Small differences in tilt, and differences in F1 bandwidth and turbulence noise suggested that some of these speakers had glottal chinks of various sizes. This group also showed a wide range of open quotients. The other four speakers (Group 2) showed substantial spectral tilts and larger open quotients compared to Group 1. The larger tilt appears to be the result of nonsimultaneous closure of the vocal folds. These speakers also showed greater amounts of noise, and, except for one, had larger F1 bandwidths than the other nine subjects, indicating larger glottal openings.

3. SPEECH ANALYSIS

Data were collected from 22 adult female speakers of American English. The materials were the three vowels /æ, Λ, ɛ/ in the /bVd/ context, embedded in the carrier phrase "Say bVd again." The subjects were instructed to speak in a normal voice. Each word was repeated five times.

The utterances were low-pass filtered at 4.5 kHz, and digitized with a sampling rate of 11.4 kHz. Spectra were obtained by first multiplying the speech signal with a 299-pt. (22.3 ms) hamming window and then computing a 512-pt. DFT. These spectra were used as the basis of the spectral measures defined above, and to estimate F1 and F3. The F1 and F3 estimates were used to set the center frequencies of the band-pass filters (BW=600 Hz) used to obtain waveforms for the waveform-based bandwidth estimates and noise ratings, respectively.

For the vowels /Λ, ɛ/, the measures H1*-H2*, H1*-A1, and H1*-A3* were taken at three points during the steady state portion of each token, for each speaker. For /æ/, we made those measurements plus the bandwidth measure and the waveform-based noise rating. The spectral-based measures for /æ/ were made at eight consecutive pitch periods during the steady state portion of each token. The bandwidth measure was made on the same pitch periods. The noise ratings were based on the F3 waveform corresponding to those eight pitch periods. The ratings were made independantly by two judges who did not know the identity of the speakers. The ratings by the two judges were well correlated and were averaged to obtain a single rating for each token. For each vowel, all measures were averaged across all tokens to obtain a mean for each speaker.

The range of values for these measures is considerable [5]. H1*-H2* has a range of about 10 dB, corresponding to a 40% range in open quotient [5], and H1*-A3* has a range of more than 25 dB, indicating a wide variation in spectral tilt among these subjects. This large range of spectral tilt is assumed to be a consequence of the presence of a glottal chink or a nonsimultaneous glottal closure along the length of the glottis, or both, for some speakers. The range of H1*-A1 is 16 dB, showing that F1 peaks vary from being very prominent for some speakers to being highly damped for others. This range of F1 amplitudes presumably arises in part due to variations of F1 bandwidths and in part due to degree to which spectral tilt extends to the low frequency harmonics.

The F1 bandwidth estimates for /æ/ vary from 41 Hz to 280 Hz. For the four speakers with the lowest values of bandwidth (up to 74 Hz), the data are in the range expected for the closed-glottis condition [7]. For speakers with higher values, losses must exist at the glottis. The noise ratings for /æ/ have a range of 1.1 to 3.8, indicating that some of our subjects show little to no noise in the high frequency range, while other speakers have substantial noise.

Correlations between the measures were computed. The highest correlation was between H1*-A3* and the noise ratings for /æ/, with $r = 0.88$. This result supports H1*-A3* as a measure of spectral tilt because a higher tilt can be caused by a glottal chink or the vocal folds being open at the vocal processes, both of which should result in greater noise generation at the glottis. We performed a regression analysis on the data, resulting in an r^2 value of 0.76.

For the vowels /ɛ/ and /Λ/, the correlation between H1*-A3* (tilt) and H1*-A1* (F1 prominence) is significant, with $r = 0.70$ and $r = 0.79$, respectively. Thus H1*-A1* is also a good indicator of the degree of glottal opening. For /æ/ the correlation is weaker, with $r = 0.63$. For the three vowels combined, $r = 0.64$. Fig. 1 is a plot of H1*-A3* vs. H1*-A1 for all vowels. The points seem to fall into two groups, with one group having tilt less than 25 dB and H1*-A1 less than -2 dB (Group 1), and the other group having tilt greater than 25 dB and H1*-A1 greater than -6 dB (Group 2). Twelve speakers fall into Group 2 for almost all of their vowel tokens. These speakers, then, have considerable spectral tilts that possibly extend to the F1 region, a significant amount of noise at high frequencies, and a strongly damped F1 peak, all evidence of a glottal opening that extends to the vocal processes [5]. The remaining 10 speakers fall into Group 1 for most tokens of their vowels. With the small amount of tilt that they exhibit, these speakers probably have closure at the vocal processes. However, members of this group with higher values of H1*-A1 also have higher tilts, suggesting that some speakers may have a glottal opening during the closed part of the glottal cycle, and the size of this opening varies with speaker.

For /ae/, the estimated bandwidth measure is not strongly correlated with H1*-A1 ($r = 0.63$), but Fig. 2 shows that there is a definite increase in H1*-A1 as bandwidth increases. In the plot four outliers are indicated. When these are removed from the analysis, the correlation increases considerably to 0.82, suggesting that for most speakers H1*-A1 may be a good predictor of F1 bandwidth.

From the acoustic analysis we can make assumptions about the glottal configurations of our subjects during phon-

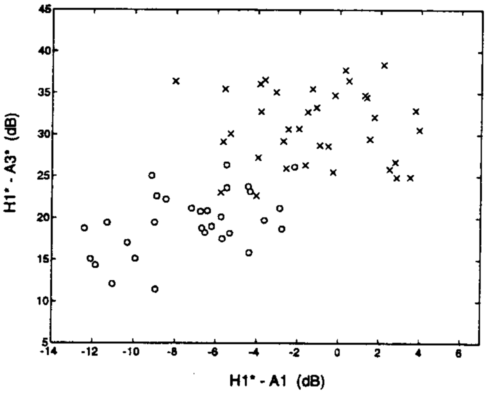

Figure 1: *Spectral tilt vs. prominence of the F1 peak for all vowels. Tokens for members of Group 1 are indicated with an* o, *while those for Group 2 are indicated with an* x.

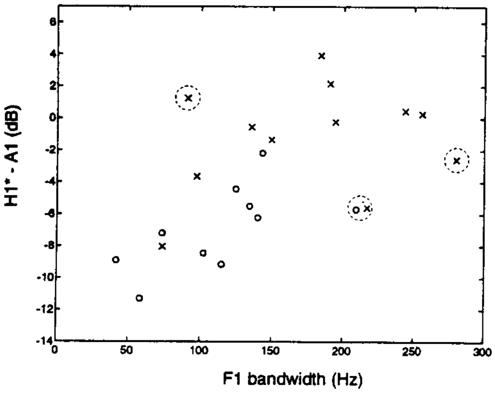

Figure 2: *Prominence of the F1 peak vs. the F1 bandwidth estimated from the speech waveform, for the vowel /æ/. The correlation coefficient is 0.63, but when outliers (indicated by broken circles) are removed, the correlation increases (r = 0.82). Tokens are again divided to indicate speakers believed to have different glottal configurations.*

ation. To verify these assumptions, we have begun to carry out some additional experiments, which we now discuss.

4. GLOTTAL AIRFLOW-BASED MEASURES

We are making aerodynamic measures on four of our speakers, using the methods of Holmberg, et al. [2]. The same utterances as in Section 3 were recorded, but in addition to the sound pressure signal, intraoral air pressure and oral airflow were recorded using a Rothenberg mask. The oral airflow signal was inverse-filtered to obtain a glottal airflow waveform. Glottal characteristics were extracted from the glottal airflow waveform and its derivative.

Data for two subjects, *hh* and *kp*, have been analyzed. Average values (across vowels) for the acoustic and airflow data of these two subjects are summarized in Table 1, along with average values of the acoustic data for all 22 subjects and average values for airflow data of a group of 15 females reported by Perkell et al. [8]. Based on the acoustic data, the two subjects fall into Group 1, theorized to have closure at the vocal processes during the glottal cycle. Both have narrow F1 bandwidths, indicating that closure may be nearly complete. From Table 1 we see that they are well below average for all acoustic measures except H1*-H2*. The glottal airflow data supports this interpretation. Minimum (DC) flow is below the group average reported in [8], suggesting smaller glottal openings (see also the ratio of AC to DC flow). Maximum flow declination rate (MFDR) is well above average, suggesting very abrupt closure, which would be unlikely if closure extended to the vocal processes.

Table 1: *Average acoustic and airflow data for two subjects,* hh *and* kp, *compared with averages over all subjects. The average airflow data are from Perkell et al. [8].*

	hh	*kp*	*Averages over all subjects*
H1*-H2* (dB)	2.4	4.8	3.1
H1*-A1 (dB)	-6.0	-11.9	-3.9
H1*-A3* (dB)	21.8	17.5	25.8
F1 bandwidth (Hz)	73	58	148
Noise rating	1.1	1.2	2.1
Peak flow (l/s)	0.22	0.28	0.29
Minimum flow (l/s)	0.06	0.10	0.13
AC flow (l/s)	0.15	0.18	0.16
Open quotient	0.42	0.55	-
Speed quotient	2.72	3.56	-
MFDR (l/s^2)	349	458	184
ACDC ratio	0.51	0.42	0.30

The airflow data analyzed thus far supports our interpretation of the acoustic data. We will also be analyzing airflow data from two subjects in Group 2 in order to confirm these results.

5. LISTENING TESTS

Results of the acoustic analysis suggest that 12 of our speakers (Group 2) have large spectral tilts, significant noise excitation at high frequencies, and strongly damped first formants. We have suggested that these measures imply glottal configurations that have large openings during the closed part of the glottal cycle, probably extending to the vocal processes. Such a configuration has been associated with breathy voice [4], [5], so if the acoustic measures are accurate, listeners should perceive Group 2 as being breathier than the speakers in Group 1, and voices should be easier to distinguish across the two groups than within the group. To test these hypotheses, we ran two listening tests, based loosely on the method of Gelfer [9].

The stimuli for the tests were vowels excised from the running speech recorded in Section 3. For both tests, one token of each vowel for each speaker was used. In the first test, listeners made similarity ratings on all possible pairs of nine speakers. Three of these speakers were chosen from Group 2 and six were chosen from Group 1. The Group 1 speakers were chosen so that three had very narrow bandwidth estimates (called Group 1a), indicating that complete glottal closure might occur during the glottal cycle. These speakers were expected to be perceived as the least breathy. The other three members (Group 1b) were chosen from among those hypothesized, based on the acoustic data, to have an intermediate level of breathiness compared to the other six speakers. The pairs were randomized and presented in both A-B and B-A order. For the second test, listeners made ratings on breathiness for all 22 subjects. For each test item, an excised vowel was repeated three times in succession. The test items were randomized and each was presented several times during the test.

There were four listeners, three female and one male, all speech researchers with experience doing listening tests. For the first test, the listeners were asked to concentrate on voice, rather than vowel quality and pitch, and to indicate how similar the voices were on a scale from "very similar" to "not at all similar." In the second test, the listeners were asked to make ratings along a scale from "breathy" to "full voice." Following the tests, the ratings were converted to scores on a scale from 0 to 7. All listeners found the tasks to be very difficult, saying that it was not easy to ignore such strong perceptual factors as vowel quality and pitch. Nevertheless, some conclusions can be made from the results.

For the first test we expected that similarity ratings would be highest when voices were compared within the three hypothesized groups and that listeners would find members of Group 2 less similar when compared to the Group 1a subjects than when compared to the Group 1b subjects. Comparisons between Group 1b and Groups 1a or 2 should result in some intermediate level of similarity. From the test results it was clear that when speakers are compared within a group, they are usally judged to be more similar than when speakers across groups are compared. In particular the Group 1 speakers with narrow bandwidths are considered more dissimilar when compared to the Group 2 speakers than when they are compared to the Group 1b speakers.

The results of the second test were averaged across vowels and listeners, and the speakers were then ranked for breathiness. Except for one speaker, all members of Group 2 were judged to be breathier than Group 1 members. The average rating for Group 1 was 5.0, while for Group 2 it was 2.6, where 0.0 corresponds to "breathy" and 7.0 to "full voice". Many researchers have suggested H1-H2 as a measure of breathiness (see [4] for a review), but, surprisingly, for our data H1*-H2* was not a good correlate of perceived breathiness. However, H1*-A1 and H1*-A3* were moderately to strongly correlated with perceived breathiness, with some variation among listeners. This is not surprising given that the division of the subjects into two groups was largely based on these two measures (see Fig. 1), and Group 2 is perceived to be breathier than Group 1.

Although these results support our method of classifying glottal configurations based on acoustic measurements, they are only preliminary. More complete testing must be done with more listeners. While it is ideal to have judgements made on natural running speech, cleaner results might be obtained using sustained vowels instead. Also, we will do tests using synthesized speech, which will allow us to carefully control glottal parameters, while eliminating the effects of vowel quality and pitch that made the tests reported here so difficult for the listeners.

6. CONCLUSION

In conclusion, we have presented measures of glottal characteristics that can be taken on the acoustic speech signal and used to classify speakers according to glottal configuration during phonation. We have shown that our interpretations of these measurements are consistent with physiological and perceptual measures. These acoustic measures are easier to obtain than physiologically-based measures and will make the use of voice source parameters in speech processing applications more practical. In addition, application of these measures to a group of female subjects has shown that more than half of these speakers show substantial spectral tilt, F1 damping, and high frequency noise, thus underscoring the importance of paying attention to such details in studies of voice and speech recognition and speech synthesis.

Acknowledgements: Many thanks to Eva Holmberg for her assistance in gathering and analyzing the airflow data, to Alice Turk for help in designing the listening tests, and to Ken Stevens for suggestions and encouragement.

7. REFERENCES

[1] H. M. Hanson. "Synthesis of female speech using the Klatt synthesizer," *Speech Communication Group Working Papers*, MIT, Cambridge, MA, to appear.

[2] E. B. Holmberg, R. E. Hillman and J. S. Perkell. "Glottal airflow and pressure measurements for soft, normal and loud voice by male and female speakers," *J. Acoust. Soc. Am.*, 84, 511-529, 1988.

[3] S. Kiritani, H. Imagawa, and H. Hirose. "Vocal cord vibration and voice source characteristics: Observations by a high-speed digital image recording," in *Proc. ICSLP-90*, Kobe, Japan, 61-64.

[4] D. Klatt and L. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, 87, 820-857, 1990.

[5] K. N. Stevens and H. M. Hanson. "Classification of glottal vibration from acoustic measurements," in O. Fujimura and M. Hirano (eds.), *Vocal Fold Physiology: Voice Quality Control*, San Diego: Singular, in press.

[6] G. Fant. "Glottal source and excitation analysis," *Speech Transmission Lab. Quarterly Progress and Status Report*, 1, Royal Institute of Technology, Stockholm, 85-107, 1979.

[7] Fant, G. "Formant bandwidth data," *Speech Transmission Lab. Quarterly Progress and Status Report*, 1, Royal Institute of Technology, Stockholm, 1-3, 1962.

[8] J. S. Perkell, R. E. Hillman, and E. B. Holmberg. "Group differences in measures of voice production and revised values of maximum airflow declination rate," *J. Acoust. Soc. Am.*, 96, 695-698, 1994.

[9] M. P. Gelfer. "A multidimensional scaling study of voice quality in females," *Phonetica*, 50, 15-27, 1993.