Analysis of Glottal Waveforms Across Stress Styles

Kathleen E. Cummings and Mark A. Clements
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

1 ABSTRACT

This paper describes results from a study of glottal waveforms derived from eleven different styles of stressed speech. The general goal of this research has been to investigate the differences in the glottal excitation across specific speaking styles. Using a glottal inverse filtering method tailored specifically for the speech under study, an analysis of glottal waveforms from voiced speech spoken under various conditions has been carried out. The extracted waveforms have been parameterized using six descriptors, and statistical analysis of the resulting database has allowed very good characterization of each style with respect to these parameters. Specifically, each style has been shown to have a unique profile which will allow high performance in a stress style identification task.

2 INTRODUCTION

Accurate modelling of the glottal source has long been a subject of great interest in speech processing. There are many applications for glottal modelling including speech coding[1], synthesis[2], and recognition[3]. The goal of this research has been to describe the characteristics of the glottal waveforms of eleven types of stressed speech and to use this knowledge to improve automatic recognition of stressed speech. This improvement could be made either by directly incorporating glottal effects into the recognition process or by identifying the speech style with its glottal waveform and choosing an appropriate recognition/compensation algorithm.

It is generally believed that one of the major conveyors of stress is the manner of glottal excitation (i.e., the glottal source waveform). If one wishes to study how stress and speaking style affect speech, it is therefore important to be able to examine the glottal source waveform. Unfortunately, most existing analysis methods do not allow for convenient separation of the glottal and vocal tract effects; hence, reliable information about changes in the glottal waveform caused by stress is difficult to obtain. Additionally, extracting the glottal waveforms from stressed speech presents specific difficulties. The most significant difficulty derives from the fact that the pitch period and other characteristics vary across speech styles. For example, in *angry* and *loud* speech both the pitch period and the interval of glottal closure are very short. Further, *question* exhibits a pitch period which rapidly changes during an utterance. The vocal tract can also be very difficult to model under certain stress conditions such as *50% tasking* and *70% tasking*. Because it is important to maintain consistency in the method of extraction over the styles of speech, all of the differences in the stressed speech must be accounted for in the design of the extraction method.

In this paper, a method for extracting the glottal waveform from a segment of voiced speech is presented. This method is based on a procedure suggested by Wong et al.[4] which has been specifically tailored for this application. The glottal waveforms for the eleven types of stressed speech contained in the Lincoln Labs Multi-Style Speech Database - ***normal, angry, loud, soft, slow, fast, clear, question, 50% tasking, 70% tasking, and Lombard*** - have been extracted from utterances of the vowel /I/ in "fix" and "six" and the vowel /ϵ/ in "destination." These glottal waveforms have been analyzed statistically and qualitatively to identify glottal characteristics that are unique to a given stress style.

3 THEORY

In the standard speech model we assume

$$S(z) = G(z)V(z)R(z), \tag{1}$$

where

$S(z)$ = z-Transform of speech waveform, $s(n)$
$G(z)$ = z-Transform of glottal source waveform, $g(n)$
$V(z)$ = vocal tract filter
$R(z)$ = radiation impedance at the lips

Therefore,

$$G(z) = \frac{S(z)}{V(z)R(z)} \tag{2}$$

Thus, $g(n)$ can be obtained by inverse filtering $s(n)$ by $v(n)$ and $r(n)$. For the voiced, non-nasalized phonemes used in this research, the vocal tract can be modelled as an all-pole filter.

Radiation impedance at the lips has often been modelled as a filter consisting of one or two zeroes. We have chosen, based on the work of Barnwell et al.[5], to model this radiation as a zero-pair and a pole-pair. In Barnwell's research, it was hypothesized that one difference between LPC synthesized speech and natural speech is a result of the glottal pulse being non-minimum phase. It was found that this can be compensated by using a fixed second-order pre-emphasis filter (i.e., two zeroes) and a ten-pole vocal tract model (for 8 kHz-sampled speech).

This material is based upon work supported under a National Science Foundation Graduate Fellowship.

CH2847-2/90/0000-0369 $1.00 © 1990 IEEE

We incorporated these results by modelling the effects of the radiation at the lips impedance as a pole-pair and a zero-pair.

Covariance LPC analysis minimizes the error, e(n), where

$$e(n) = s(n) + \Sigma_k a(k)s(n-k). \quad (3)$$

If the coefficients, a(k), model the vocal tract perfectly and the radiation impedance at the lips has been compensated, e(n) would equal the glottal source, g(n). The estimation of the vocal tract parameters is most accurate during the period of glottal closure, since at this time, s(n) is theoretically a freely decaying oscillation affected only by the vocal tract and radiation at the lips. Therefore, an all-pole model for s(n) is a good assumption during glottal closure and produces a relatively accurate model of v(n). In this procedure, covariance LPC analysis is performed over windows no longer than the expected period of glottal closure of s(n), shifted by one sample at a time. When the error waveform corresponding to this analysis is examined, glottal closure is assumed to exist over the segment for which e(n) is close to zero. Since the vocal tiact has constraints on how rapidly it can change, the model found during the identified glottal closure is used to inverse filter an interval long enough to include four pitch periods of s(n) - the output being four periods of g(n).

4 METHOD

In order to compare and draw conclusions about changes in the glottal waveform caused by stress and speaking style, it is important that the extractions be performed under the same conditions in each case. The actual modelling of the vocal tract must be done over the same portion of a glottal period to minimize differences between the glottal waveforms of different speaking styles caused by differences in the modelling technique. This poses a particular problem because of the variability of the pitch period for different types of stressed speech. Such styles as ***angry*** and ***loud*** have pitch periods which are so short that the vocal tract can not be accurately modelled with the samples comprising glottal closure. There is, therefore, a trade-off between using the ideal segment - glottal closure - and maintaining consistency across the speaking styles by using the same length segment of a glottal period for each extraction. In order to maximize the number of samples in the vocal tract model, to maintain the accuracy of the model, and to maintain consistency across the stress styles, the extractions were performed over segments slightly longer than the apparent glottal closure.

The first step in the extraction of the glottal source waveform from a given utterance is computing e(n) using covariance LPC analysis. The LPC analysis is performed over a segment of windowed speech using a ten pole filter (four pole-pairs for the vocal tract and one pole-pair for radiation at the lips). This window is shifted by one sample at a time and e(n) is computed for each window of modelled speech. The error waveform is then examined and a starting point is chosen where e(n) is close to zero for a number of samples. The vocal tract is again modelled at this point, and this model is used to inverse filter four pitch periods of the speech waveform, s(n). This result is integrated twice to account for the previously described zero-pair in radiation at the lips to produce the glottal source waveform, g(n). The analysis window is then shifted by one sample, the vocal tract is modelled again, and the inverse filtering and integration are repeated to produce a new g(n). This is done five to ten times and the best g(n) is chosen. In general, the extracted waveforms are extremely similar in shape and form. These iterations are carried out in order to find the best g(n) for those stress styles which are more difficult to model (e.g., ***50% tasking***, ***70% tasking***, and ***Lombard***). After careful study, we elected to describe a period of a glottal waveform by six parameters. These are: opening slope, closing slope, opening duration, top duration, closing duration, and closed duration. These parameters are illustrated in Figure 1. The four duration parameters are measured as the number of samples between the endpoints and including one endpoint. The two slope parameters are measured as the slope at the onset of opening and the offset of closing.

At this point, the amplitude of g(n) is dependent only on the error, e(n), at the point where the inverse filter was computed. This error can be affected by many things. In order to make accurate comparisons of the areas and slopes of the various styles of glottal waveforms, all of the amplitudes were normalized with respect to ***normal***. First, all glottal waveforms are normalized to have the same maximum amplitude. Each waveform is then multiplied by a factor of (given stress intensity)/(***normal*** intensity) and divided by (given stress open-closed ratio)/(***normal*** open-closed ratio). The intensity figures used are from previous research[6] and the open-closed ratios (ratio of open part of glottal pulse to closed part of glottal pulse) used are from this research. In this way, the amplitudes compared to normal are meaningful and comparisons between the styles are valid.

This process was performed on 6-8 utterances for each of the eleven stress styles resulting in 50-100 pitch periods of glottal source for each style of speech. Both the /I/ and /ϵ/ contexts were used to ensure that the results were not vowel specific. Furthermore, the glottal waveforms from a second speaker were examined to ensure that the results were not speaker dependent. Each pitch period of glottal source was hand-marked to extract the six parameters, producing 66 distributions - one for every stress style for each parameter.

Several statistical tests were performed on these distributions. The Kolmogorov Smirnoff (K-S) test, which is a nonparametric test that compares two distributions to decide if the distributions are not significantly different, was used in two ways. First, the K-S test was performed on each pairwise comparison of the 66 distributions. Secondly, the K-S test was used to compare each of the 66 distributions to its respective best-fit Gaussian distribution. This was done in order to ascertain whether the assumption that the distributions were Gaussian was a fair assumption. In order to confirm the results of the second set of K-S tests, a Chi-Square goodness-of-fit test was also performed, comparing each of the 66 distributions to its best-fit Gaussian. On the basis of the results of these two tests, the best-fit Gaussian mean and variance were computed for each of the 66 distributions.

5 RESULTS

Example glottal waveforms for each of the eleven speaking styles are presented in Figure 2. The K-S test versus a Gaussian and the Chi-Square goodness-of-fit test for a Gaussian both showed that the assumption that these distributions are Gaussian is not a bad one. The best-fit Gaussian means

	Closing Slope	Opening Slope	Closed Dur	Closing Dur	Opening Dur	Top Dur
Angry	-9910	9198	9.1	6.3	6.9	2.0
50%	-4522	2321	17.3	11.1	16.0	9.8
Clear	-5011	2686	15.8	9.5	16.0	6.9
70%	-4100	2138	16.7	10.7	15.7	9.9
Fast	-3972	2376	15.5	11.0	16.0	8.4
Loud	-9298	3532	6.3	6.9	17.0	2.9
Lombard	-5430	2871	15.2	9.3	15.2	7.6
Normal	-4798	2643	17.7	10.2	15.6	9.9
Question	-4831	3034	14.0	9.4	14.9	7.0
Slow	-4786	2692	16.9	10.2	15.5	8.7
Soft	-2632	1921	17.7	14.7	18.6	9.9

Table 1. Best fit Gaussian means.

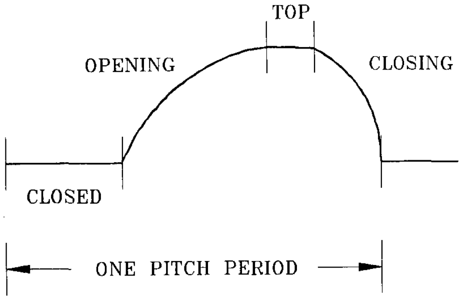


Figure 1: Example marked period of a glottal waveform.

are provided in Table 1. The most interesting results obtained from examining the statistics in Table 1 are presented in Table 2. An important result is that, with the exception of *question*, the percentage of the pitch period during which the glottis is closed is fairly constant across the speaking styles. More significant are the variations across the speaking styles of the ratios of closing to opening slope and closing to opening duration. These ratios are different enough that they may be used to identify a speaking style. Considering the ratio of closing to opening slope, ***angry***, ***fast***, ***loud***, ***question***, and ***soft*** are significantly different from each of the other speaking styles. Similarly, ***angry***, ***loud***, and ***soft*** have quite different closing to opening duration ratios.

The results from the pairwise K-S tests are important for the task of identifying a speaking style with its glottal waveform if a Bayesian classifier is used. For this to work the distributions must be significantly different for each speaking style using a linear combination of the six parameters. The results from the K-S pairwise tests show that the parameters which contain the most information about a speaking style are closing slope, opening slope, and closed duration. This is as expected since the slopes of glottal opening and closing provide most of the control over the amount of acoustic energy produced. Furthermore, since the duration of glottal closure is approximately a constant percentage of the pitch period for a given speaking style, the closed duration parameter directly reflects the length of the pitch period. It is well accepted that one important conveyor of stress is the length of the pitch period. Further examination of statistics from the tests indicate that each speaking style is significantly different from each of the other ten using combinations of the six parameters.

	Closing vs. Opening Slope	Closing vs. Opening Duration	% Pitch Period Closed
Angry	1.1	.92	31
50%	2.0	.69	30
Clear	1.9	.60	30
70%	1.9	.68	30
Fast	1.7	.69	29
Loud	2.6	.40	23
Lombard	1.9	.60	31
Normal	1.8	.65	31
Question	1.6	.63	36
Slow	1.8	.66	30
Soft	1.4	.79	30

Table 2. Results generated using best fit Gaussian means.

6 CONCLUSION

The results of this research show that each of the eleven styles of glottal waveforms has a unique profile based on the six parameters: closing slope, opening slope, opening duration, top duration, closing duration, and closed duration. This is not a profile which depends on the vowel spoken. Furthermore, examination of the glottal waveforms of a second speaker suggests that the results are not speaker dependent. Once the glottal extraction procedure has been automated, these statistical profiles can be used to identify the style of a given utterance of speech from a voiced segment of that utterance. Once the speaking style is known, automatic recognition of the speech can be improved by choosing a specifically designed enhancement algorithm or a specifically trained codebook. These results should allow for significant improvement of automatic recognition of stressed and otherwise style-variant speech. Another possibly useful application would be that of changing one speaking style into another by LPC synthesis using appropriately modified excitation waveforms.

References

[1] A. Bergstrom, and P. Hedelin, "Code Book Driven Glottal Pulse Analysis," *Proceedings, 1989 International Conference on Acoustics, Speech, and Signal Processing*, vol. S1, pp 53-56, May, 1989.

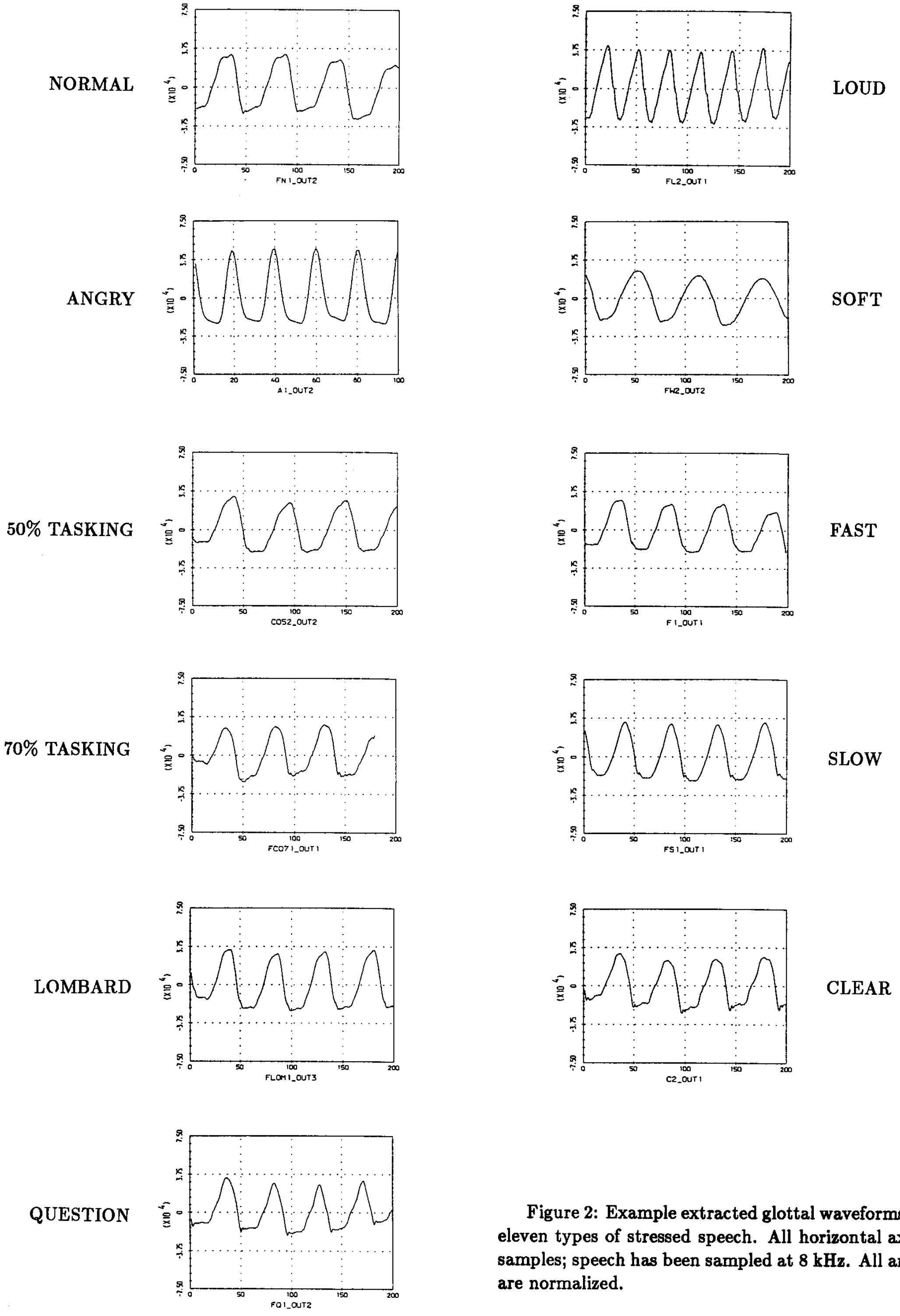
[2] R. Carlson, G. Fant, C. Gobl, I. Karlsson, and Q. Lin, "Voice Source Rules for Text-to-Speech Synthesis," *Proceedings, 1989 International Conference on Acoustics, Speech, and Signal Processing*, vol. S1, pp 223-226, May, 1989.

[3] J. Hansen and M. Clements, "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition," *Proceedings, 1989 International Conference on Acoustics, Speech, and Signal Processing*, vol. S1, pp 266-269, May, 1989.

[4] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 4, pp. 350-355, August, 1979.

[5] T. P. Barnwell, R. W. Schafer, and A. M. Bush, "Tandem Interconnections of LPC and CVSD Digital Speech Coders," *Technical Report E21-685-77-TB-2, Goergia Institute of Technology*, November, 1977.

[6] J. H. L. Hansen, *Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition*, Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, July, 1988.


NORMAL

LOUD

ANGRY

SOFT

50% TASKING

FAST

70% TASKING

SLOW

LOMBARD

CLEAR

QUESTION

Figure 2: Example extracted glottal waveforms from the eleven types of stressed speech. All horizontal axes are in samples; speech has been sampled at 8 kHz. All amplitudes are normalized.