IMPROVEMENTS TO AND APPLICATIONS OF ANALYSIS OF STRESSED SPEECH USING GLOTTAL WAVEFORMS

Kathleen E. Cummings and Mark A. Clements

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

1. ABSTRACT

This paper presents a continuation of previous research in the area of statistical analysis and modelling of glottal waveforms of emotionally stressed speech [1]. A simple glottal model will be defined which, by varying four parameters, can be used to model the glottal excitation of many styles of speech. An algorithm which automatically decouples these model glottal waveforms from the vocal tract information in a given speech segment will also be presented. Several applications of decoupling the model glottal waveforms will be discussed. In particular, a method of automatic formant tracking of stressed speech which is superior to traditional formant tracking procedures will be described. Also, a method of modifying the style of speech by removing some of the effects of stress on the speech waveform will be presented.

2. INTRODUCTION

Modelling the glottal excitation waveform has long been an area of great interest in speech processing. An accurate and simple glottal model would have important applications in areas such as speech coding, speech synthesis, and automatic recognition. Another area which has received considerable attention in recent years is the analysis of speech produced under varying styles and levels of emotional stress. Understanding the components of speech that are correlated with the style of speech has applications in many areas including speech synthesis and recognition.

In previously reported work [1], the combination of these two problems was addressed. "Pseudo-glottal" waveforms (henceforth, the term "glottal waveform" will be used to signify this) were extracted from eleven styles of speech in the Lincoln Labs Style Database using an interactive glottal inverse filtering method. These styles include *normal, angry, loud, soft, fast, slow, 50% tasking, 70% tasking, Lombard, question,* and *clear.* Several utterances of three different vowels from three different speakers were used. Once extracted, each pitch period of the glottal waveforms was hand-marked into four segments: closed, opening, top, and closing. Several parameters including the waveform slopes at the onset of opening and closing, duration of opening and closing, and duration of closure were analyzed statistically. It was determined that these parameters were significantly different across speech styles and that these differences were consistent across utterances, vowels, and speakers.

This paper will present a continuation of this previous research and will focus on two developments that provide the groundwork for applications. First, a simple glottal model which can accurately represent the salient features in virtually all glottal waveshapes will be defined. Second, an algorithm which decouples these glottal waveforms from the vocal tract signal will be described. Finally, several of the many possible applications of these developments will be presented. Automatic formant tracking and speech style modification will be discussed in some detail. The applications of stress style identification and constrained automatic glottal modelling will also be described.

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

3. THEORY

In the standard source-filter theory of speech production [2], voiced speech is assumed to be the result of convolving a quasi-periodic train of impulses with short-time stationary filters representing glottal shaping, vocal tract resonances, and radiation impedance at the lips. In terms of z–transforms, this model can be expressed as

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

representing, respectively, the z–transforms of the speech signal, $s(n)$, the impulse train, $e(n)$, the glottal pulse, $g(n)$, the vocal tract impulse response, $v(n)$, and the radiation at the lips, $r(n)$.

The production of a single pitch period of voiced speech, denoted $s_i(n)$, (hereafter, the subscript i should be understood to mean the ith pitch period) can be modelled as a single impulse convolved with a glottal excitation pulse and the impulse responses of the vocal tract and radiation at the lips. The small decaying speech signal from the previous pitch period, which is additive in time, is usually either considered negligible or accounted for with initial conditions. In terms of z–transforms,

$$S_i(z) = 1 \cdot G_i(z)V_i(z)R(z). \tag{2}$$

Given $s_i(n)$ and assuming that radiation can be modelled as a zero at or near $z = 1$ ($R(z) = 1\text{-}\alpha z^{-1}$), the ith pitch period of speech with the radiation effects removed, $\hat{S}_i(z)$, can be represented as

$$\hat{S}_i(z) = \frac{S_i(z)}{1 - \alpha z^{-1}} = G_i(z)V_i(z), \tag{3}$$

where α is approximately 1 [3].

From this it is clear that, given a pitch period of speech with the radiation effects filtered out and an accurate glottal model, the vocal tract and glottal excitation signals can approximately be decoupled from each other as

$$\frac{\hat{S}_i(z)}{G_i(z)} = V_i(z). \tag{4}$$

0-7803-0532-9/92 $3.00

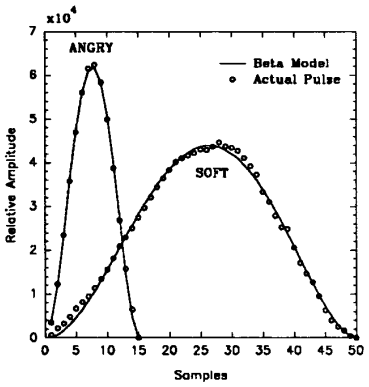

Figure 1: *Angry* and *soft* beta function models and actual extracted pulses.

Ideally, the accuracy of $V_i(z)$ is a function only of the accuracy of the glottal model, $G_i(z)$ and of the decoupling procedure.

4. CURRENT EXTENSIONS

In this section, the two components of this research that form the basis for applications of modelling the glottal waveforms of stressed speech are described. First, a simple model glottal waveform that can represent all eleven styles of glottal waveforms will be described. The reasons for choosing this model and the actual models for the eleven styles will be given. Second, a parametric algorithm for decoupling the model glottal excitation and vocal tract signals on a pitch period by pitch period basis will be presented.

4.1. Modelling Glottal Excitation

Many pulse shapes have been suggested to model the glottal excitation waveform. Among the more popular and successful of them are the polynomial and raised cosine models developed by Rosenberg, et al. [4], and the trigonometric model suggested by Hedelin, et al. [5].

Modelling the glottal waveforms of differing styles of speech, however, poses particular problems. The model must be able to describe widely variant opening and closing slopes, durations, and amount of symmetry. Furthermore, in order for the model to be useful for decoupling applications, it must be simple to use and able to describe the salient glottal waveshape parameters. In previous research [1], it was determined that the most statistically significant parameters are opening slope, closing slope, the ratio of opening to closing slope (amount of symmetry), and pulse duration. Because of its characteristics and simplicity, the beta function,

$$y = A\left(\frac{x}{x_{max}}\right)^b \left(1 - \frac{x}{x_{max}}\right)^c, 0 < x \leq x_{max}, \quad (5)$$

was chosen as a glottal model. Two example beta function models, along with the extracted glottal pulses, for the styles *angry* and *soft* are shown in Figure 1. The amount of symmetry in the pulse shape and the relative steepness of the slopes are determined by the values of b and c. Unlike many pulse definitions, the beta function is able to model slopes as slow as those found in *soft* and slopes as steep as those found in *angry* and *loud*. By varying A, b, c, and x_{max}, many glottal waveshapes, including the eleven analyzed styles, can be modelled accurately.

Style	A	b	c	x_{max}	% RMS
normal	249218	1.388	.924	31	6.3
angry	983219	2.045	1.946	14	1.9
loud	534016	2.236	1.100	29	12.8
soft	683960	2.095	1.876	49	3.4
fast	209372	1.336	1.070	33	4.6
slow	311293	1.734	1.049	36	4.2
50% task	280533	1.609	1.028	33	4.6
70% task	279266	1.749	1.098	34	4.6
Lombard	323321	1.716	1.131	29	7.1
question	263356	1.665	0.968	33	6.4
clear	386780	1.813	1.127	34	7.0

Table 1: Beta function parameters for modelling styled glottal waveforms.

For each of the eleven styles of extracted glottal waveform, a median pulse from the distributions of extracted glottal pulses was selected and the beta function which minimized the least squared error to this glottal pulse was found. The parameters A, b, c, and x_{max} and the RMS error as a percentage of the RMS value of the original pulse are shown in Table 1. For most of the styles, the RMS error was around 5% of the RMS pulse value. Our observation is that most of this error is caused in fitting the top of the pulse shape, the least significant of the parameters of the extracted waveforms. The error in the fit of the *loud* glottal pulse, at 12.8%, was by far the worst.

4.2. Decoupling Algorithm

Successful decoupling of the vocal tract and the model glottal excitation signals has two immediate results. First, the accuracy of a glottal model can be verified by removing the glottal signal and determining whether the remaining signal is indeed representative of the vocal tract. Second, synthesizing styled speech, stress style identification based on glottal waveshape, and automatic glottal modelling, all of which depend on isolating either the vocal tract or the glottal excitation signal, can be implemented.

The algorithm developed is based on the model glottal waveform described above and is performed on a pitch period by pitch period basis. A data file was developed for each of the eleven styles of speech containing the following parameters: A, b, c, and average % of the pitch period in which the glottis is open (this is essentially pulse duration as a % of pitch period), denoted $T_{\%open}$. The pitch contour and the boundary samples were found for each segment of voiced speech to be analyzed. These are stored in vectors of the form $[T_1, T_2, \ldots, T_i, \ldots T_N]$, where T_i is the number of samples in the ith pitch period, and $[V_{beg}, V_{end}]$, where the values are the beginning and ending samples of voicing for the utterance as a whole. Except for marking the boundaries of the voiced segment and checking the automatic pitch detection, the algorithm is automatic.

The analysis is begun by filtering out the effects of radiation, leaving $\hat{s}(n)$. The following steps are then followed to determine a reconstructed vocal tract impulse response, $v(n)$:

1. Select the first analysis frame, $\hat{s}_1(n)$, as the speech segment extending from V_{beg} to $V_{beg} + T_1 - 1$. Multiply by a Hamming window of length T_1. Set i = 1.
2. Calculate $x_{max_i} = T_{\%open} \cdot T_i$.

3. Define model glottal waveform $g_i(n)$ in time using x_{max_i} and A, b, and c from the datafile.

4. Take FFTs of $\hat{s}_i(n)$ and $g_i(n)$.

5. Divide $\hat{S}_i(e^{j\omega})$ by $G_i(e^{j\omega})$ to get $V_i(e^{j\omega})$.

6. Take IFFT of $V_i(e^{j\omega})$ and multiply by the inverse of a Hamming window to get $v_i(n)$.

7. Construct $v(n)$ by lining up the first sample of $v_i(n)$ with the first sample of $\hat{s}_i(n)$. The decaying signal will extend beyond the actual pitch period to the length of the FFT. The overlap in adjacent $v_i(n)$'s is added together.

8. Set $i = i+1$. Select $\hat{s}_i(n)$ as the speech segment beginning one sample past the end of the previous analysis window and extending T_i - 1 samples.

9. Follow steps 2 – 8 until the end of the voiced segment is reached.

5. APPLICATIONS

Given styled model glottal waveforms and an algorithm to decouple the vocal tract and model glottal signals, there are several possible applications. Two have been implemented and will be discussed in some detail in this paper. Two more, owing to the preliminary nature of their implementations, will be discussed more briefly.

The first implementation was developed as a method of checking the accuracy of the glottal modelling and decoupling. It is, however, an important application in itself. This is automatic formant tracking of stressed speech using the residual vocal tract signal, $v(n)$. The second is the application of resynthesizing speech in order to modify the perceived style or stress level of the original speech. The latter two, which are being developed currently, are automatic stress style identification and constrained automatic glottal modelling.

5.1. Automatic Formant Tracking of Stressed Speech

Determining the accuracy of the glottal model and the decoupling procedure is not an easy task. One method of testing the consistency of the model is to examine the linear prediction (LP) formant tracks of the residual signal, $v(n)$. If the modelling and decoupling were performed correctly, $v(n)$ should be well-modelled by 8 poles. The formant tracks should be smooth and at appropriate frequencies. If, on the other hand, either the modelling or the decoupling are in error, the 8-pole LP model would be expected to have difficulty tracking the formants smoothly and might have significant formant frequency errors.

Voiced segments from the utterances "go" and "eight" from one speaker and "fix" and "help" from another were analyzed, and the LP formant tracks of the resulting $v(n)$'s were determined. Note that these voiced segments include a diphthong and a semivowel in addition to vowels. With the single exception of the style *angry*, all of the formant tracks were smooth and were located at appropriate frequencies. This strongly suggests that both the glottal models and the decoupling procedure are accurate. An example of the formant tracks from a *loud* utterance of "help" are shown in the bottom part of Figure 2. These formant tracks are compared to those found using a standard formant tracking method and will be discussed in more detail later in this section.

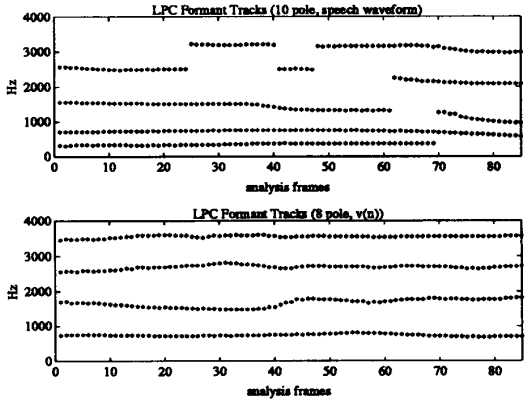

Figure 2: LP formant tracks of the original speech and of $v(n)$, respectively.

Automatic formant tracking of stressed speech using the glottal models and the decoupling algorithm is actually an important application. Although automatic formant tracking of voiced speech is generally considered a solved problem, in the case of stressed speech it can still be a difficult task. Because the spectral content of some styles of glottal waveforms can vary significantly, performing 10-pole analysis of the speech and attributing the lowest amplitude or real poles to the glottal spectrum may not be accurate. Sometimes the lowest amplitude poles actually correspond to the third or fourth formant, and occasionally the poles may be shifted slightly in frequency to model low frequency glottal energy. An example comparing automatic formant tracks from the /ϵ/–/l/ in "help" , spoken in the *loud* condition, can be seen in Figure 2. The first track was found using standard 10-pole analysis in which the four complex pole-pairs with the highest magnitudes are identified as representing the four formants. The second was found using an 8-pole analysis of $v(n)$. The second analysis resulted in excellent formant tracks, where the more standard method had significant difficulty selecting the formants. In fact, the 8-pole LP analysis of $v(n)$ almost always results in more accurate formant tracking than does the standard method for all of the styles of speech in the Lincoln Labs Style database.

It is clear for stressed speech that performing automatic formant tracking on $v(n)$ is far superior to the standard method. Additionally, decoupling of $v(n)$ from the excitation allows for higher order modelling of the vocal tract to be performed easily because there is no need to determine which poles are modelling the vocal tract and which are modelling the glottal excitation.

In the case of *angry*, which has always proven to be a difficult style of speech to analyze, the formants occasionally were found to lie on the real axis at $\omega = \pi$. Previously, it has been observed [6] that *angry* speech has lower energy in low frequencies and higher energy in high frequencies than other styles of speech. In examining the power spectra of both $s(n)$ and $v(n)$, this observation was confirmed. It is felt that the LP formant locations on the real axis are a result of this high frequency energy concentration and perhaps the need for a higher sampling frequency and not of the method or modelling.

5.2. Speech Style Modification

Modifying the style of a given speech segment by re-exciting $v(n)$ with a different style model glottal waveform is one

of the more important applications of the previous developments. Resynthesizing styled speech as *normal* speech is in essence removing some of the effects of stress in the speech. Performing this modification prior to automatic speech recognition should improve automatic recognition accuracy. Alternatively, *normal* speech could be made to sound emotionally stressed.

This is in practice much more difficult than the simple theory suggests. In addition to glottal waveshape, other acoustic correlates such as average phoneme duration and average pitch must be accounted for. It is known that, with the exception of *question*, the shape of the pitch contour of a given utterance tends to be maintained across speech styles. The average pitch, however, does vary significantly with changes in speech style [6]. Therefore, the new speech is resynthesized with the same pitch contour but with an appropriately valued average pitch. Furthermore, the ideal duration of the new utterance is calculated using the ratio of the statistical means of the durations of the new to the original style of speech. An optimal arrangement for repeating or deleting pitch periods as the new speech is concatenated is used to achieve this ideal duration.

The algorithm to modify the style of speech is a fairly straightforward extension of the decoupling algorithm. Rather than concatenating $v(n)$, pitch periods of new speech are produced and are concatenated into a new style of the original voiced utterance. Thus far, this has been implemented two different ways: with an FFT-based method and with a Linear Predictive model-based method.

The FFT-based implementation is the most straightforward extension of the decoupling algorithm possible. After the original speech spectrum is divided by the original model glottal spectrum, the resulting vocal tract spectrum is multiplied by the new model glottal spectrum. The IFFT of the resulting signal is taken to find a pitch period of new speech. The new speech is concatenated using the same overlap-add method described in Step #8 of the decoupling algorithm.

In the LP model-based implementation, $v(n)$ is concatenated as described in the decoupling algorithm. Linear predictive analysis with 8 poles is then performed on windows of $v(n)$, incrementing by exactly one pitch period length along the entire duration of $v(n)$. The new model glottal excitation is convolved in time with the impulse response of each pitch period's LP model to produce a pitch period of new speech. These pitch periods are concatenated into a new version of the original utterance using the overlap-add method described in Step #8 of the decoupling algorithm.

Because the speech style modification results are perceptual in nature, they cannot be discussed quantitatively. In listening to the stressed speech that has been resynthesized as *normal* speech, it is clear that all of the styles sound closer to *normal*. Conversely, the effects of adding stress to *normal* speech, that is, resynthesizing *normal* speech as styled speech, can also be heard. In particular, changing *normal* speech to *Lombard*, *fast*, or *slow* speech is quite distinct.

5.3. Other Applications

Two other applications of the styled glottal models and the decoupling algorithm are being actively researched. Constrained automatic glottal modelling, in which the optimal beta function glottal model is found, is the first of these applications. An iterative procedure to find the best model by removing the beta function models from speech and selecting the "best" vocal tract is being developed. The second of the two applications is the task of stress style identification of a given speech token. Two methods are being considered to implement this task. The first is to remove all eleven styles of glottal models from a speech segment and examine the LP error of modelling the resulting $v(n)$'s. The second method is to convolve each of the eleven styled glottal waveforms with a "good" vocal tract signal and determine which is most likely to have produced the speech.

6. CONCLUSION

The problems of modelling the glottal excitation of stressed speech and of developing applications of this modelling have been considered in this paper. Based on previous statistical analysis defining unique glottal pulse shape parameters, a simple model which is able to define all of the salient glottal shape features easily was found in the beta function. The actual beta functions for the eleven styles of stressed glottal waveforms considered were determined. An algorithm was developed to decouple this model glottal waveform from the vocal tract signal in a given segment of voiced speech.

The glottal model and the decoupling algorithm were shown to be accurate by examining the LP formant tracks of the remaining vocal tract signal. Several applications, including automatic formant tracking and modifying the speech style, were discussed. Automatic formant tracking of stressed speech using the glottal model and the decoupling algorithm was shown to be superior to normal methods. Two implementations of re-exciting the vocal tract signal with a new model glottal waveform were presented. Finally, the two tasks of stress style identification and constrained automatic glottal modelling were discussed.

The results of this work are very promising. By automatically identifying the stress style of a given segment of speech and modifying it to be more *normal*, it should be possible to improve automatic recognition rates of stressed speech.

REFERENCES

[1] Cummings, K. E. and Clements, M. A. "Analysis of Glottal Waveforms Across Stress Styles," *Proceedings, IEEE ICASSP*, 1990, pp 369-372.

[2] Fant, C. G., *Acoustic Theory of Speech Production*, The Hague: Mouton, pp 15-62, 1970.

[3] Markel, J. D. and Gray, Jr. A. H., *Linear Prediciton of Speech*, New York: Springer–Verlag, pg 7, 1976.

[4] Rosenberg, A. E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *The Journal of the Acoustical Society of America*, Vol 49, pp 583-590, 1971.

[5] Hedelin, P., "High Quality Glottal LPC-Vocoding," *Proceedings, IEEE ICASSP*, 1986, pp 465-468.

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