Proceedings of The Institute of Acoustics

SPECTRAL PROPERTIES OF GLOTTAL FLOW PULSES AS A FUNCTION OF SPEAKERS, VOWEL AND STRESS LEVEL

L. BOVES

INSTITUTE OF PHONETICS, NIJMEGEN UNIVERSITY, THE NETHERLANDS

INTRODUCTION

This paper attempts to answer three questions that can be asked regarding the spectra of the glottal volume velocity pulses reconstructed from actual speech:
- Can speakers be separated on the basis of the volume flow spectra?
- Can the pulses pertaining to stressed vowels be separated from those pertaining to unstressed vowels?
- Are the pulse shapes vowel dependent?

INVERSE FILTERING

It is well known that it is possible to recover the glottal flow by passing the speech signal through a filter that is the exact inverse of the filter formed by the combined nasal and vocal tracts. In order to insure the stability of the inverse filter the procedure is mostly confined to speech sounds for which the filter can be assumed to be all-pole. This, of course, leaves only non-nasalised vowels as candidates for processing.
The derivation of the all-pole model of acoustic vowel production relies on the assumption that the acoustic tube formed by the vocal tract is closed at the glottis. This assumption clearly does not hold during the open glottis interval. Model simulations, inspection of oscillograms and actual formant measurements carried out on very short signal segments show that the parameters of the filter during the closed and open glottis intervals are considerably different. These differences must be attributed to the changing termination impedance at the glottis. Consequently, the parameters of the inverse filter should be estimated from the portions of the speech signal pertaining to the closed glottis interval, since it is only there that the all-pole model is a good approximation.
Wong et al. [1] have argued that the closed glottis interval can be established by locating local minima in the normalised prediction error obtained from a sequential covariance analysis using a very short rectangular window. Our experience, however, shows that the normalised prediction error is highly dependent on the formant pattern of the vowel under analysis [2, 3]. Therefore, we prefer the electroglottogram, recorded simultaneously with the speech signal, as a means of determining the moment of glottal closure. In our analysis system the parameters of the inverse filter are estimated in terms of the filter coefficients obtained from a covariance LP analysis carried out on a signal segment immediately following the moment of glottal closure. Window length is usually between 24 and 40 samples and the predictor order is 8 or 10. Since we intend to process running speech, an estimate of the parameters is obtained for each pitch period and the inverse filter is updated pitch-synchronously.

SPECTRAL PROPERTIES OF GLOTTAL FLOW PULSES AS A FUNCTION OF SPEAKERS, VOWEL AND STRESS LEVEL

SPEECH MATERIAL

Four adult male speakers, selected for their auditorily quite differing voice qualities, read aloud a neutral text that took about 2 minutes to complete. Speech signal and electroglottogram were directly digitised and stored on the computer's disc. Although this procedure insures a minimal amount of low frequency phase distortion, phase correction was nevertheless necessary to undo the effect of the 22.2 Hz high pass filter in the microphone preamplifier. The signals were sampled at a rate of 8 kHz per channel.
From each reading 60 vowels were selected subject to the conditions that they should be surrounded by pauses or by non-nasal consonants and that about half of the vowels should come from stressed syllables. The vowels taken from unstressed syllables should not be reduced to such an extent as to lose their identity. For each speaker a number of schwa vowels were, however, included deliberately. The neutral character of the text did not give rise to very emphatic reading, so that the differences between stressed and unstressed vowels are not overly apparent. We think, however, that they are fairly representative of normal unemotional reading.

SIGNAL AND DATA PROCESSING

From all 240 vowels the glottal volume flow waveform was reconstructed using the technique indicated above. This was done interactively, i.e. the output was monitored, and the window length and predictor order were adjusted to give a result that was maximally consistent with the open quotient estimated from the electroglottogram. A portion of the flow waveform from the most central part of the vowel was isolated by multiplying a Hamming window into the flow signal. The length of the window was adjusted so that an integral number of periods was contained within in. The windowed flow signal was padded with zeros to obtain a record of 1024 samples that was subjected to an FFT. The constant bandwidth FFT spectra were next converted into 'critical band' spectra by summing the appropriate number of adjacent Fourier coefficients. Thirteen pass-bands suffice to cover the range from 80 Hz up to 4 kHz. The filter levels were expressed in dB relative to the overall level determined by summing all 512 Fourier coefficients.
After these operations the volume flow waveforms are reduced to points in a 13-dimensional space. Experience with critical band spectra of speech signals has invariably shown that the number of dimensions can be reduced considerably without losing much information. From the many possible techniques we have chosen multiple discriminant analysis [4]. Essentially this is a mapping of a source space into a destination space under the constraint that groups of data points should occupy maximally different regions in the destination space. The dimensions spanning the destination space can be interpreted in a number of ways, e.g. by the relative contributions of the dimensions of the source space and by careful study of the group centroids in the destination space.

RESULTS

The speaker separation experiment resulted in 75% of the glottal pulse spectra being assigned to the correct group. Of the three discriminant functions the first and by far the most important one seemed to be primarily related to f_o. The two speakers with f_o around 100 Hz were clearly separated from the remaining

SPECTRAL PROPERTIES OF GLOTTAL FLOW PULSES AS A FUNCTION OF SPEAKERS, VOWEL AND STRESS LEVEL

speakers, whose f_o is about 140 Hz. The second and third discriminant functions were very hard to interpret but seemed to be most important for the separation of the individuals in the high and low pitch groups. The 75% accuracy in the speaker classification is disappointingly low, especially with a view to the fact that the speakers were selected for their differing voice qualities. It might be that these differences are mainly due to dynamic aspects of the glottal pulses, which are no longer present in our data.

The attempt to separate the stressed and unstressed vowels was carried out for the four speakers separately. The percentages of the spectra classified correctly ranged from 62% (which is only just beyond chance) to 83%. This corresponds well with the auditory impression of some readings being slightly more emphatic than others. Only for the two most emphatic speakers (i.e. those who yielded the highest percentages of correcly classified spectra) the a priori assumption was confirmed that stressed vowels would be characterised by less steeply falling slopes in the glottal source spectra. Mean spectra, spectral variation and spectral differences between stressed and unstressed vowel pulses are illustrated in Fig. 1.

For each of the speakers it was attempted to separate the pulses pertaining to three groups of vowels, viz. the high vowels /u, y, i/, the mid vowels /e, I, ɛ, o, ɔ, ʌ/ and the low vowels /a, ɑ/. The stressed and unstressed vowels were taken together; the schwa vowels were treated as low vowels mainly to get groups containing about the same number of elements. The percentages of spectra classified correctly ranged from 63% to 70%. For all speakers almost all confusions occurred between low and mid vowels, whereas the high vowels appeared to form a clearly separate category. The differences between the group centroids in the original 13-dimensional space are depicted in Fig.2, as an aid in interpreting the results. From that figure it seems that our data support Rothenberg's contention [5] that the spectra of the pulses pertaining to high vowels are characterised by steeper

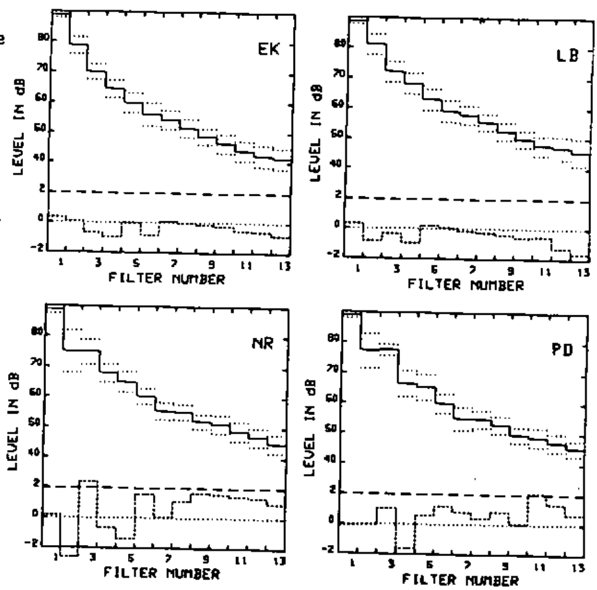

Fig. 1 Average critical bandwidth spectra of the flow pulses of four speakers (upper parts of the panels) and spectral differences between pulses pertaining to stressed and unstressed vowels (lower parts of the panels). Dotted lines in the upper panels indicate $\pm$ 1σ regions around the mean.

slopes in the frequency region around 600 Hz than those pertaining to other vowels. The effect may be explained by the fact that the vowel tract forms an inductive load for frequencies below F1. For the high vowels this load is confined to a much more narrow frequency range than for the remaining vowels.

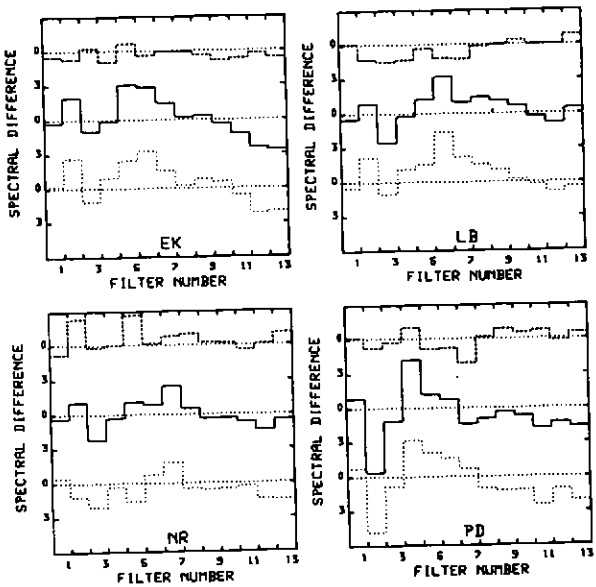

Fig. 2 Spectral differences of mean spectra of the flow pulses reconstructed from three groups of vowels.
Dashed lines : low vowels - mid vowels
Continuous lines : low vowels - high vowels
Dotted lines : mid vowels - high vowels

REFERENCES

[1] D.Y. WONG, J. MARKEL and A.H. GRAY 1979 IEEE Trans.Acoust.,Speech,Signal Processing, ASSP-27, 350-355.
Least squares glottal inverse filtering from the acoustic speech waveform.

[2] L. BOVES and B. CRANEN 1982 Conf.Rec. ICASSP '82, Paris, 1988-1991.
Evaluation of glottal inverse filtering by means of physiological registrations.

[3] L. BOVES and B. CRANEN 1982 in: Fortschritte der Akustik, FASE/DAGA '82, Göttingen, Berlin: VDE-Verlag.
Comparison of reconstructed glottal flow waveforms with physiological registrations.

[4] J. OVERALL and C. KLATT 1972 New York: McGraw Hill.
Applied multivariate analysis.

[5] M. ROTHENBERG 1982 STL-QPRS, 1, 1-17.
An interactive model for the voice source.