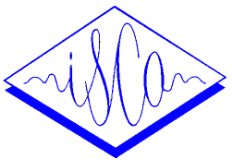

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

COMPARISON OF INVERSE FILTERING OF THE FLOW SIGNAL AND MICROPHONE SIGNAL.

**R Orr, *B. Cranen, **F. de Jong and *L. Boves

*Department of Voice and Speech, ENT Section, St. Radboud University Hospital, NL-6500 HB, Nijmegen
** Department of Language and Speech, Nijmegen University, NL-6500 HD Nijmegen

ABSTRACT

This study looks at two ways of extracting a glottal waveform from recorded speech. One way is to inverse filter the flow at the mouth. Another is to inverse filter the microphone signal. Theoretically, the microphone signal is considered to be the equivalent of a first order differentiation of the flow signal recorded at the lips.

Recording the oral airflow is more complicated than the recording of a microphone signal, as it requires the use of a mask, with constant adjustments during the recording. Recording of the microphone signal is more straightforward for the experimenter and less intrusive for the subject. If the two inverse filtering procedures can be shown to produce similar glottal flow waveforms for both types of recorded speech, this would support the use of only the microphone signal for those types of glottal flow analysis where the DC component of the flow is not essetial, making voice source analysis applicable in less specialised situations.

In this study, we used recordings of microphone signal and recordings of oral flow to compare the results of inverse filtering. A group of twenty subjects produced repetitions of the utterance /pae/. We recorded oral flow, EGG, and the microphone signal. The flow and microphone signals were analysed using an automatic inverse filtering program and values for parameters which are extracted from the source wave are compared.

The results were not as similar as expected, although in some respects, they correlated well. This may be due to experimental design, the degree of insight of the subject into the voicing task, and the fact that the speech material used for the comparison was not identical.

1. INTRODUCTION

Inverse filtering is now a well-established method for examination of the voice source (1, 2). It may be described simply as a technique in which the effects of the formants created by the vocal tract filter are removed from the speech signal, leaving only the flow signal as the excitation of the vocal tract. Two types of speech signals are commonly inverse filtered in order to arrive at a representation of the voice source. The first of these is the airflow recorded at the mouth. The second is the microphone signal.

The recording of speech with a pressure sensitive microphone is essentially a transformation of the volume-velocity airflow to air pressure at a distance from the lips. The effect of this transformation is modelled as a boosting of the spectrum by 6dB per octave, corresponding to a first order derivative of the signal. It is therefore generally assumed that the microphone signal is equivalent to a first order derivative of the airflow at the mouth.

The flow signal has two major advantages over the microphone signal. The first is that absolute values for DC leakage can be measured if the signal is accurately calibrated. The presence of, and amount of DC leakage is an important aspect of glottal behaviour. The second is that the simplified assumption above, regarding the effects of lip radiation does not have to be considered.

However, there are disadvantages with measuring the flow signal. The signal is generally recorded with the aid of a flow-mask (3) which does not pick up high frequency components in the source signal, thus limiting the amount of available information. The use of the mask itself may be a hindrance for some subjects, which may have an undesired effect on the data. With some subject groups, for example children, it may be entirely unsuitable. These problems do not have to be present if the microphone signal can be used.

It is the aim of this paper to investigate whether glottal flow estimates obtained from the microphone signal are sufficiently similar to estimates derived from the flow signal to substitute the latter with the former, at least for those research and clinical diagnostic goals where the DC component of the flow is not essential. In this way, we hope to lend more credibility to the use of the microphone signal in inverse filtering, where this is the experimenter's chosen method.

2. EXPERIMENTAL SETUP

2.1 Subjects

The subjects were a group of 20 students of teaching, who were participating in a larger project, investigating suitability of the voice for the teaching profession. All subjects underwent logopedic and phoniatric examination, and were only admitted to the group if there were no voice problems and if the vocal folds were healthy. The group consisted of 16 women and 6 men, with an age range of 18 to 36.

2.2 Phonation Task

Each subject was asked to produce at least four /pæ/ syllables in five different voicing conditions, namely, high pitch, low pitch, high intensity (loud voice), low intensity (soft voice), and normal pitch and intensity. For each /pæpæpæpæ/ utterance, the subject was asked to sustain the last /pæ/ for a few seconds.

2.3 Measurements

The above phonation tasks were recorded under four different experimental conditions, with various recording combinations of microphone signal, flow signal, EGG signal, oral pressure, and videolaryngostroboscopy. The combiation of signals which was actually measured is given in Table 1 below.

Condition	Microphone	Flow signal	Oral Pressure	EGG	Videolaryn go-
1	▭			▭	
2	▭			▭	▭
3		▭	▭	▭	
4		▭	▭	▭	▭

Table 1: measurements taken per experimental condition

Of course, it would be ideal to be able to use identical speech for the experimental material, but the microphone signal recording of speech through the face mask was not considered acceptable for inverse filtering.

2.4 Recording Procedure

The microphone recordings were made using a Bruel and Kjaer microphone (4133) and a Bruel and Kjaer amplifier (2619). The oral flow and pressure were measured with a circumferentially vented pneumotachograph mask (Glottal Enterprises) with a heated double screen wire mesh, in combination with a Glottal Enterprises amplifier (MS-100A2). Directly before and after the flow recordings, both the flow and pressure sensors were calibrated in order to get absolute pressure and flow measures.

The EGG signals were recorded with a (Laryngograph Ltd.) laryngograph. The videostroboscopic images were recorded on a Super-VHS videorecorder using a flexible endoscope (Olympus ENF type P3) and a Kay RLS 9100 Rhino-Laryngeal Stroboscope. The speech recordings made while the subject was undergoing stroboscopy are included as data in the analysis in this paper. This data was collected as part of the aforementioned larger project on the professional teaching voice, and it is assumed here that phonation produced under endoscopic intrusion constitutes a separate voicing condition. For this reason, a description of the recordings is provided.

The signals were recorded on a 14-channel FM-recorder (TEAC XR510). The recordings were made at a tape speed of 19.05 cm/s. For optimal use of the available dynamic range, the microphone signals were recorded on three different channels with low, medium and high input gains. The EGG and flow signals were similarly recorded, each at two different levels, on two different channels.

2.5 Signal Processing

All signals were digitised at a 10 kHz sampling rate. The EGG recording was used to determine the time of closurefor each period, using a peak-picking algorithm on the differentiated glottogram signal. The microphone and flow signals were then automatically inverse filtered using covariance LPC on the closed glottis interval.

2.6 Parameterisation of the source

Parameters from the time domain, like open quotient (OQ), closed quotient (CQ), and speed quotient (SQ) depend on determination of moments of opening and closing. The often gradual nature of opening of the vocal folds makes the definition of an actual moment of opening somewhat difficult to define with any confidence. In order to avoid possible uncertainty, time domain parameters were not included in this analysis.

The inverse filtered waveforms were visually inspected, and any sections where the waveform was unsatisfactorily filtered were removed. The remaining signal was divided into equal length sections of 1024 samples. This was done in order to obtain some idea of the inter-subject variability. From this data, we calculated some of the spectral parameters similar to those used in other research (for example, 4, 5) **.** The spectral peaks below 1500 Hz were then assumed to be harmonics, and their frequencies and amplitudes were recorded. The spectral parameters used in this study are summarised in Table 2.

Parameter	*Description*
F0	the frequency of the first harmonic

H1-H2 (dB)	difference between the amplitudes of the first and second harmonics
spectral slope (dB/oct)	regression line calculated from the amplitudes of the harmonics below 1500Hz

Table 2: parameters used for the comparison of inverse filtering methods.

3. RESULTS

The two types of speech signal, microphone and flow, could not be recorded simultaneously, as the flow mask rendered any microphone signal useless for inverse filtering. Therefore, we recorded independent productions of /pæpæ/ utterances. Although the recordings were made during one session, there remains the difficulty of using non-identical recordings in order to compare two measurement methods. While the results cannot be expected to be identical, we should at least expect good correlations between similar instances of speech filtered using the different methods.

For each parameter, we looked at the means of the samples per person , and we compare d these means per voicing condition, for experiments 1 and 3, and for experiments 2 and 4. For each calculation of a mean value, there were at least 20 samples. This was to ensure that the values were representative of the entire utterance, and that the filtering was successful for that utterance. We then used a correlation test (Pearson Product Moment) to examine the correlation between the inverse filtering methods

Comparison of experimental conditions 1(mic) and 3(flow)			
voicing condition	**F0**	**H1-H2**	**slope dB/oct**
loud	0.59	0.24	0.81
low	0.75	0.004	0.3
normal	0.91	0.07	0.1
soft	0.14	0.02	0.24

Table 3: p values for comparison of means (paired ttests) of F0, H1-H2 and the spectral slope for the voicing conditions loud, low (pitch), normal and soft (intensity). The comparison is between experimental condition 2(microphone recording) and condition 4 (flow recordings).

Comparison of experimental conditions 2(mic) and 4(flow)			
voicing condition	**F0**	**H1-H2**	**slope dB/oct**
loud	0.34	0.32	0.3
low	0.62	0.37	0.03
normal	0.23	0.12	0.01
soft	0.06	0.18	0.45

Table 4: p values for comparison of means (paired ttests) of F0, H1-H2 and the spectral slope for the voicing conditions loud, low (pitch), normal and soft (intensity). The comparison is between experimental condition 2(microphone recording with endoscopic intrusion) and condition 4 (flow recordings with endoscopic intrusion).

Comparison of experimental conditions 1(mic) and 3(flow)			
voicing condition	**F0**	**H1-H2**	**slope dB/oct**
loud	0.64	0.75	0.54
low	0.92	0.51	0.68
normal	0.95	0.34	0.63
soft	0.98	0.31	0.34

Table 5: correlation values (r) for mean values of F0, H1-H2 and the spectral slope for the voicing conditions loud, low (pitch), normal and soft (intensity). The correlation measurement is between experimental condition 1(microphone recording) and condition 3 (flow recordings)..

Comparison of experimental conditions 2(mic) and 4(flow)			
voicing condition	**F0**	**H1-H2**	**slope dB/oct**
loud	0.93	-0.48	0.11
low	0.7	0.51	0.5
normal	0.9	0.3	0.73
soft	0.97	0.4	0.64

Table 6*:* correlation values (r) for mean values of F0, H1-H2 and the spectral slope for the voicing conditions loud, low (pitch), normal and soft (intensity). The correlation measurement is between experimental condition 2(microphone recording with endoscopic intrusion) and condition 4 (flow recordings with endoscopic intrusion).

Tables 3 and 4 show the p-values for paired t-tests and tables 5 and 6 show the r-values for correlation tests (Pearson Product Moment) carried out on the spectral data. We looked at the data in two ways as neither gives a complete picture. Even if a *t*-test does not show that the parameter values differ significantly between the two measurement techniques, we would still want to know how well the two sets of measurements coincide. The value of the co-efficient *r* gives at least some indication in this respect

4. DISCUSSION

4.1 *t*-tests

T-tests were unable to find any significant differences between the two sets of data. This is only to be expected, as the method of filtering should not affect the period time. However, significant differences were found between the two measurement methods for some of the spectral parameters, namely the conditions of low and soft voice for the H1-H2 parameter, and conditions of normal and low voice for spectral slope.

Where there is a low p-value, such as 0.004, for the comparison of H1-H2 means for low voice, we might be inclined to conclude that the flow mask systematically interferes with the subjects' speech to a significant extent. However, this cannot be concluded for all of the data.

4.2 Correlation Tests

From the results, it is clear that the fundamental frequencies were mostly comparable between the two measurement conditions, with correlation values generally above r=0.9. What is surprising is that the correlation is not higher. In particluar, the values of r= 0.64 for loud voice, in the comparison of experiments 1and 3, and of r=0.7 for low voice in the comparison of experiments 2 and 4 are poor.

The same method was used for all data to find the fundamental period per pulse, that is, peak picking in the EGG signal in order to find moments of closure. Furthermore, the closure markers were all subsequently visually checked for accuracy. It is unlikely that the inconsistencies lie here. Intra-subject variability in consecutive /pæpæ/ productions in a single session was not high. The obvious explanation of variation in F0 is that the measurements were carried out on non-identical speech samples. However, values as low as r=0.64 cannot thus be accounted for.

It is more likely that the subject did not always interpret the experimenter's instructions in exactly the same way. For example, it is possible that the instruction to speak in a low voice may have been interpreted as low intensity for one experiment and as low pitch for another. The production of loud voice may have been accompanied by different degrees of increased pitch. This explanation is supported by subsequently listening to the original recordings.

The spectral parameters were chosen for this investigation because of their robust nature. They are not reliant on features of the glottal pulse which are difficult to measure. Nor are they highly susceptible to measurement artifacts, like slight errors in phase correction. They are also commonly used as indicators of voice quality, and therfore relevant to this area of research. We would have strong expectations that they would correlate well for the two measurement methods.

The results show that these expectations were not entirely borne out. Better correlation was observed for spectral slope than for H1-H2. This is particularly evident for the normal voicing condition. The fundamental frequency with which the subjects - mostly female - phonated was generally around 220-230Hz. The second harmonic was therefore likely to occur around the same frequency as the first formant for the /æ/ vowel, causing boosting of the harmonic amplitude. Small variation in F0 may have produced much bigger variation in H2 levels. This may account for the lower correlation values found for this parameter.

We did, however, find some correlation between data with and without the mask. For example, H1-H2 measurements correlate well for loud voice in experiments 1 and 3, and spectral slope measurments correlate well for normal voice in experiments 2 and 4.

A possible explanation for the generally low correlation values for the spectral parameters lies in the nature of the design of this study. Both H1-H2 and spectral slope reflect glottal behaviour strategies. A soft voice can be produced with increased glottal opening, which corresponds to an increased spectral slope, and a more dominant H2. Soft voice may also be produced by reducing the subglottal pressure, without extra glottal opening. This would not necessarily produce a steeper slope. Rather, the speaker may lower their F0, and the voice might become creaky which could even increase the slope. Similar observations may be made about the production of loud voice and low voice. Although source parameters, by design, should be sensitive to different voicing conditions, inclusion of parameters from the time and intensity domains might have yielded better overall results.

There is no reason to assume that speakers are consistent in the strategies which they use. There is also no reason to assume that subjects in this type of research have any great insight into how they produce different voicing conditions. It is difficult to conceive of an experimental setup where the protocol could incorporate such strict control of the subjects that different voicing conditions would be produced in a consistent manner. Even were this possible, it is doubtful that the resulting data would be representative of natural speech.

5. CONCLUSIONS

In this paper, we investigated the results of inverse filtering of speech from two different types of signal. It was expected that parameters extracted from the different types of filtering would be comparable. The nature of the measurements required that the data came from different recordings. We hoped, with a strict data collection protocol, to elicit speech samples which would be similar enough for such a comparison.

The results indicate that parameter values extracted from the data are not always highly correlated. However, there is no pattern evident which suggests a systematic difference in the filtering methods. Indications from the fundamental frequency results are that the data sets are not sufficiently similar for this

comparison. Indications from the spectral parameters are that they are too sensitive to inconsistent voicing strategies on the part of the subject.

We suggest that not only may one speaker employ a number of different methods to vary the voicing conditon, but that the experimenter may be faced with misinterpretation of his or her instructions. It seems that, in order to compare two different analysis methods, using non-identical speech samples, a protocol would have to be devised where subjects would be so tightly controlled that the resulting speech would probably not be representative of the speaker's natural voice.

For this type of comparison, it seems that simultaneously recorded data must be used. This might require a different method of collecting the flow data. A more extensive choice of parameters may produce better results.

6. REFERENCES

1. Cranen, B. "Simultaneous modelling of EGG, PGG and glottal flow." *J. Acoust. Soc. Am.*, 84:88. 1988
2. Gobl, C. and Ní Chasaide, A. "Acoustic characteristics of voice quality." *Speech Comm.* 11, 481:490. 1992
3. Rothenberg, M. "A new inverse filtering technique for deriving the glottal air flow waveform during voicing". *J. Ac. Soc. Am.* 53, 1632:1645
4. Maddieson, I. and Ladefoged, P. " 'Tense' and 'Lax' in four minority languages of China." *UCLA Working Papers in Phonetics.* 60, 59:83
5. Jackson, M., Ladefoged, P., Huffman, M .K., and Antoñanzas-Barroso, N. "Automated Measures of Spectral Tilt". *UCLA Working Papers in Phonetics.* 62, 77-88.