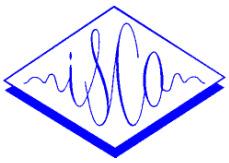
ISCA Archive
http://www.isca-speech.org/archive



2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

THE INFLUENCE OF LINGUISTIC VARIATIONS ON THE VOICE SOURCE CHARACTERISTICS

Jacques Koreman, Louis Boves & Bert Cranen

University of Nijmegen, Dept. of Language and Speech, Phonetics section
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
Phone: +31 80 615584 / E-mail: koreman@lett.kun.nl

Abstract

We will discuss a method to compare the voice source characteristics of consonants belonging to four different *manner* classes. Results from an experiment will be discussed, and the different voice source characteristics will be related to *sonority*.

1 Introduction

Although F_0 and amplitude of the glottal airflow have long been the only voice source properties that aroused interest, the importance of other voice source characteristics which determine the *shape* of the glottal pulses is now widely recognized. These characteristics include open quotient, skewing, leak flow, excitation strength, etc.

Dynamic voice source models have been developed which are able to reflect the variability in the source characteristics observed in natural speech (esp. [1, 2, 3]). So far, comparatively little systematic work has been done on the analysis of natural voice source signals to supply the models with control parameters, especially where linguistic variations are concerned, although a number of articles *have* been published in the last few years ([4, 5, 6, 7]).

In this article, we will present the results of an experiment which shows the effect of different *manners* of articulation on the voice source characteristics. We distinguish consonants belonging to four manner classes: sonorants (SON), fully voiced obstruents (VD), voiced obstruents with interruption of vocal fold vibration before the release (VDI), and voiceless obstruents (VL).

2 Method

2.1 Subjects and stimuli

We asked five male subjects (all non-smokers with no known voice disorders) to pronounce a set of carrier phrases containing a nonword of the type /pɛpɛCɛpɛ/, where *C* can be any Dutch consonant (voiced or voiceless plosive or fricative, nasal, liquid, glide or /h/). The carrier phrases were pronounced with a rise + fall intonation pattern (described by [8], as two linked H^*L's, where the linking causes the first *L* to be lost). By varying the position of the nonwords in the carrier phrase, four intonation conditions were created: one in which there is a falling pitch movement on the third syllable of the nonword (F3), one in which it is on the second (F2), and two more with a rising instead of a falling pitch movement (R3 and R2). The differences between the intonation conditions are not discussed in this article.

2.2 Deriving voice source characteristics

We recorded oral flow with a Rothenberg mask, digitised (10 kHz) and inverse filtered this signal, and then parameterised the resulting glottal flow signal. The parameters that are computed are used as descriptors of the voice source characteristics.

Our inverse filtering procedure uses the EGG as a help signal ([9]). It makes a pitch-synchronous LPC estimate of the vocal tract resonances on the basis of the closed glottis interval of each glottal pulse, the inverse of which is used to filter the oral flow signal. This results in a fairly reliable estimate of the glottal flow for each separate nonword.

There are two reasons why we do not use the glottal flow signal itself to search for systematic patterns in the voice source characteristics. First, no pattern recognition system can successfully distinguish the raw glottal flow signals for different experimental conditions. Second, the dimensions that are used to represent the glottal flow are likely to have a strong influence on the successfulness of any pattern recognition procedure. By using dimensions which have a clear physiological interpretation, we hope to have better chances of finding interpretable, systematic patterns in the voice source characteristics for different experimental conditions. We will therefore aim to obtain a number of parameters from the glottal flow signal that together can describe the most important properties of the glottal pulses. We derive the following set of parameters (PARSET) from the glottal flow signal:

- F_0: Fundamental frequency of vocal fold vibration, derived as f_s/T_0. T_0 is defined as the distance between two subsequent minima in the time derivative of the glottal flow (these minima coincide with closure of the vocal folds).
- OQ: Open quotient, i.e. the ratio of the open interval to the total period duration (T_0). The open interval (time interval for which the vocal folds are apart) is calculated as the distance from the moment at which the glottal flow exceeds the noise level (opening moment) up to the closure.
- r_k: Skewing measure which is the inverse of speed quotient. r_k is the ratio of closing to opening interval of the vocal folds. In the glottal flow signal, it is computed as the distance in samples between maximum in glottal flow and closure (closing interval), divided by the distance from the open-

10.21437/ICSLP.1992-37

ing moment to the maximum in the glottal flow (opening interval).

E_e: Excitation strength: value of the flow derivative at the closure moment.

Each glottal period is characterised by one value for each of the parameters. If there is no voicing in the signal (e.g. during voiceless obstruents), zeroes are inserted; these zeroes do not have a meaningful quantitative interpretation (see Fig. 1).

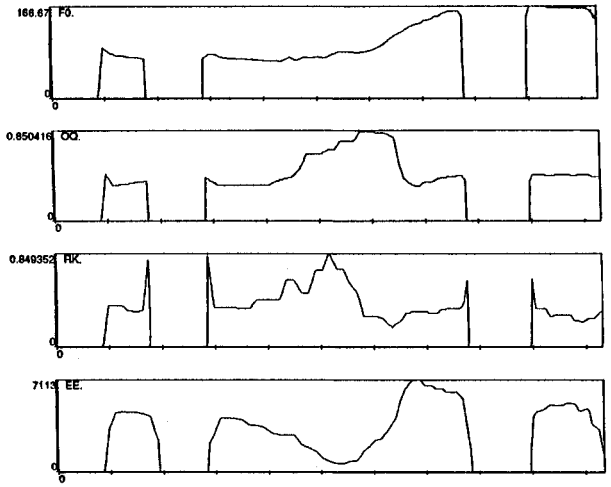


Figure 1: Parameter tracks for F_0, OQ, r_k, and E_e for /pepezepe/ (intonation condition R3)

The parameter tracks are normalised by translating the raw parameter values into z–scores[1], and multiplying them by 1000 and then adding 500[2]. Finally, all pitch synchronous parameter values were converted to uniformly sampled tracks by means of linear interpolation (parameter values were computed once every 5 ms).

The parameterisation replaces the glottal flow signal for each nonword by a set of time-aligned parameter tracks (PARSET) for each nonword. The PARSETs of all nonwords that belong to the same experimental condition must somehow be averaged to create only *one* PARSET pattern that is typical for the nonwords of that condition.

Because the zero values do not have a meaningful quantitative interpretation (i.e. zero is not 0 Hz, but indicates voicelessness), we must be careful not to average "real" values with the raw zeroes. That is why we divided the underlyingly voiced obstruents into VD and VDI: if we had not made this division, the parameter values during the voiced obstruent would be an average of the "real" parameter values during the obstruent and the inserted zero values which occur when vocal fold vibration is interrupted during closure. This might result in parameter values which can not be interpreted as possible values. For instance, a state may have an F_0 value of 40 Hz by averaging of F_0 values of 100 Hz with zero values.

The division of the underlyingly voiced obstruents into VD and VDI is made on the basis of the F_0 tracks: if zero values have been inserted for the C, i.e. if the vocal folds stop vibrating during the oral constriction for the obstruent in the onset of the third syllable, the stimulus is classified with VDI, otherwise with VD.

2.3 Finding typical patterns

In [10], we explained how Hidden Markov Modelling (HMM) can be used to create models that show typical patterns for the voice source characteristics (i.e. PARSETs) of each experimental condition, and how we can evaluate how well the models distinguish between the experimental conditions that are used in an experiment.

2.3.1 How to find patterns

One HMM is created (by a linear initial estimate + Viterbi reestimation[3]) on the basis of the PARSETs of all nonwords belonging to the same experimental condition. Each HMM consists of 16 states (which can be considered as time intervals) with typical values for each of the parameters (the parameter values are obtained as Gaussian distributions; the state durations and their standard standard deviations can be derived from the transition probabilities in the HMM).

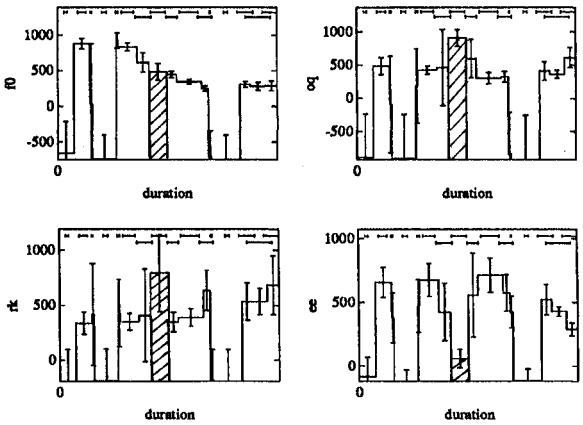


Figure 2: HMM consisting of tracks of F_0, OQ, r_k, and E_e for nonwords of the form /pepeCepe/, where C (hatched) belongs to condition VD (intonation condition F3)

Fig. 2 shows the HMM for condition VD (nonwords containing a fully voiced obstruent) of the experiment (intonation condition F3). Each nonword belonging to this condition is characterized by a PARSET of four parameters, namely F_0, OQ, r_k, and E_e. Since the HMM is optimised for the four parameter tracks simultaneously, the

[1] For the computation of the variance, the zero values were not used.

[2] This was done to prevent numeric underflow.

[3] We made use of the APMS software packet.

result shows equally long durations (with the same standard deviations) for the states of each of the four parameter tracks. For each state, the average value (and standard deviation) is computed for each of the four parameters[4].

2.3.2 How typical are the patterns?

To evaluate how different the PARSETs that belong to different experimental conditions (SON, VD, VDI, and VL) are, we will indicate *recognition results*. These recognition results, or *success rates*, are obtained in a (Viterbi) recognition procedure. Let us explain this for intonation condition F3, on the basis of the nonword /pɐpɐzɐpɐ/ (where /z/ is fully voiced). The PARSET of this nonword is first used for buiding the HMM for condition VD (together with the PARSETs of all other nonwords containing a fully voiced obstruent in the onset of the third syllable). In the experiment, models are built for conditions SON, VDI, and VL in the same manner. In the recognition procedure, for each of the four HMMs (SON, VD, VDI, and VL) the probability is computed that it generated the PARSET of /pɐpɐzɐpɐ/. If this probability is greatest for the VD HMM, the nonword is said to have been *correctly* categorised. The success rate is calculated by tallying up all correctly categorised nonwords, and dividing this number by the total number of nonwords that were used in the experiment. A high success rate means that the HMMs in the experiment are *distinctive* for the experimental conditions; a low success rate, on the other hand, means that the HMMs are quite similar, so that they cannot be considered typical of the experimental condition they model.

The recognition procedure (using the HMMs for conditions SON, VD, VDI, and VL) was carried out independently for four subexperiments, which differed in the intonation pattern which was used in realising the stimulus word (F3, F2, R3, and R2). The results from these subexperiments were very similar, so that the differences can be considered quite robust. In Table 1, the overall categorisation results are shown.

For each of the four subexperiments the percentage of correctly categorised words (success rate) is between 80 and 95%. We can conclude, therefore, that the four classes SON, VD, VDI, and VL are characterized by clearly distinguishable glottal flow pulse shapes. The percentage of correct categorisations is higher in the conditions in which the consonant is part of a stressed (F3: 92.13; R3: 94.44) than of an unstressed (F2: 80.00; R2: 85.39) syllable, which leads us to conjecture that differences in glottal articulation are maximized in stressed syllables. This tendency may be related to "articulatory effort", which has been shown to be greater in stressed syllables ([11]).

Most categorisation errors concern the mixing up the VDI and VL classes, or miscategorisations of SON nonwords.

[4]The standard deviations for the durations of the states are indicated in the upper part of each window by horizontal bars. The standard deviation for the state durations can be found by drawing an imaginary line from the middle of a state to the middle of a bar in the upper part of the window. The standard deviation of the parameter value in a state is indicated by a vertical bar in the middle of that state, centred at the average value.

Table 1: Categorisation results for SON, VD, VDI, and VL nonwords with HMMs of these four classes (summed over intonation conditions F3, F2, R3, and R2)

	SON	VD	VDI	VL
SON	87	7	6	5
VD	-	38	4	-
VDI	-	-	70	7
VL	-	-	14	120

Success rate:	88%
Number of errors:	43
Number of nonwords:	358

The mixing up of the VDI and VL classes can be easily understood. After all, the VDI and VL obstruents at the beginning of the third syllable have a lot in common: they both have a state for which all parameters are zero (though there may be a general difference in the duration of this state), and the differences that do exist between the two experimental conditions are restricted to a very short time span directly before and/or after the silent interval. The resemblance between these two classes is therefore relatively strong.

Sonorants are miscategorised as VD, VDI, or VL roughly equally often; these miscategorisations occur most frequently when the sonorants occur at the beginning of an unstressed syllable (which is immediately preceded by a stressed one in our nonwords). Sonorants therefore seem to sometimes miss their characteristic qualities at the beginning of an unstressed syllable, which allows them to be mixed up with obstruents.

None of the obstruents is ever (in any of the subexperiments) miscategorised as SON, nor is a VDI or VL stimulus ever miscategorised as VD. This is probably due to the fact that the low values which occur for the parameters during the VL or VDI obstruents when voicing is absent cannot be generated by the Hidden Markov Models for the SON or VD classes, at least not in the state(s) which describe(s) the *C*. Miscategorisation in the other direction is possible, be it at a rather high cost, by making the state for VDI or VL that describes the voiceless part of the consonant under investigation very short; the parameter values generated for the vowel preceding the VDI or VL consonant are then used as a descriptor of the VD or SON consonant.

2.4 Interpretation of the HMMs

Since the models are very distinctive (see previous section), they are visually compared, attempting to interpret them in physiological and/or linguistic terms.

Visual inspection of the Hidden Markov Models shows that E_e is lower for all consonants (SON, VD, VDI, or VL) than for the surrounding vowels, though the decrease is greater for VD (and VDI and VL, for which zero values are assigned to the excitation strength when the vocal folds stop vibrating) than for SON consonants. In the state which shows a decrease in E_e, OQ increases. For

the obstruents (VD, VDI, and VL), we notice that the increase of OQ corresponds with an increase in r_k (the shape of the glottal flow pulses becomes more sinusoidal); for the sonorants, this correspondency seems to be absent, and no striking change occurs in the value of r_k. All these findings are more tentative for the VL and VDI classes: because the changes in the parameter values occur over relatively short durations (directly before and after the voiceless interval of the consonant), HMM often does not model them by assigning a separate state to them.

3 Discussion

All the changes in the parameter values that we found for the consonants affect the energy in the part of the spectrum up to around 3 kHz: a decrease in E_e lowers the overall spectral energy, while an increase of OQ and r_k make the spectral slope steeper ([5, 12]).

In [13] the energy in the lower part of the spectrum is used for the automatic detection of syllables. In order to find syllable nuclei, amplitude peaks are detected in the microphone signal, after low–pass filtering at 2–3 kHz. The low–pass filtering suppresses the energy in the higher frequencies, as for /s/. Without low–pass filtering, the high friction energy of this sound might cause it to be considered as a syllable peak.

The energy in the lower part of the spectrum is thought of by [14] (though not uncontroversially) as an operational definition of "sonority" (pp. 56–57), which is compared to *loudness*, remarking that the definition of sonority as *loudness* has to be modified so that the loudness caused by friction is excluded.

The changes in the parameter values that we found for the consonants seem to affect the sonority of a sound. The order of sonority that we find coincides with the sonority scales that are used in phonological descriptions: obstruents have a low excitation strength, a high open quotient, and a symmetric glottal pulse shape (high r_k) when we compare them to the surrounding vowels; sonorants have a low excitation strength (though the decrease is less than for obstruents), and a high open quotient, while their skewing is the same as for vowels. This would cause the obstruents to be least sonorous, followed by sonorants, which are in turn less sonorous than vowels. The changes in the parameter values that we find suggest that the differences on the sonority scale are to a large extent caused by differences in the glottal articulation. These differences may be enhanced by the supraglottal articulation of the sounds.

References

[1] K. Ishizaka & J. Flanagan. "Synthesis of voiced sounds from a two–mass model of the vocal cords." *The Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[2] G. Fant, J. Liljencrants, & Q. Lin "A four- parameter model of glottal flow." *STL–QPSR*, vol. 4/1985, pp. 1–13, 1986.

[3] D. Klatt & L. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *JASA*, vol. 87, nr. 2, pp. 820-857, 1990.

[4] C. Gobl. "II. Speech production." *STL–QPSR*, vol. 1, pp. 123–159, 1988.

[5] C. Gobl & A. Ní Chasaide. "The effects of adjacent voiced/voiceless consonants on the vowel voice source: A cross language study". *STL–QPSR*, vol. 2–3, pp. 23–59, 1988.

[6] J. Pierrehumbert. "A preliminary study of the consequences of intonation for the voice source." *STL–QPSR*, vol. 4, pp. 23–36, 1989.

[7] A. Löfqvist & R.S. McGowan. "Influence of consonantal environment on voice source dynamics." *J. of Phon.*, vol. 20, pp. 93–110, 1992.

[8] C. Gussenhoven. "Adequacy in intonation analysis: The case of Dutch" *In:* H. van der Hulst & N. Smith (eds.). *Autosegmental studies on pitch accent*, pp. 95–121. Dordrecht: Foris, 1988.

[9] J. Koreman & B. Cranen. "(Semi–)automatic pitch–synchronous computation of glottal flow. *JASA*, vol. 86, suppl. 1, p. S36, 1989.

[10] J. Koreman, B. Cranen & L. Boves. "Automatic computation and comparison of dynamically varying voice source parameters". *Proc. of the second Eur. Conf. on Speech Comm. and Techn. (Eurospeech 91)*, vol. 3, pp. 1077–1080, 1991.

[11] A. Ní Chasaide. "Glottal Control of aspiration and voicelessness". *Proc. of the eleventh Int. Congr. of Phon. Sc.*, vol. 6, pp. 28–31, 1987.

[12] G. Fant & Q. Lin. "Frequency–domain interpretation and derivation of glottal flow parameters." *STL–QPSR*, vol. 2–3, pp. 1–21, 1988.

[13] T. Rietveld & L. Boves. "La détection automatique de syllabes accentuées en néerlandais" *Actes de IXe Journees d'Etudes sur la Parole*, Lannion, pp. 262–269, 1978.

[14] T. Rietveld. *Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands*, PhD dissertation University of Nijmegen (The Netherlands), 1984.