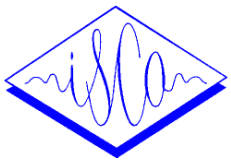

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

LINGUISTIC AND PARALINGUISTIC VARIATION IN THE VOICE SOURCE

Ailbhe Ní Chasaide and Christer Gobl***

*Centre for Language and Communication Studies, Trinity College,
University of Dublin, Ireland. **Department of Speech Communication
and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

This paper presents an overview of past results and ongoing work by the authors on voice source variation. The LF-model is used to quantify voice source parameters and ultimately as the basis for resynthesis. First of all, the nature and directionality of coarticulatory effects of voiced/voiceless segments are looked at across languages. Results show some striking cross-language differences, and some of these can be closely linked to the temporal coordination of laryngeal and supralaryngeal gestures. Voice source variations due to prosodic context (degree of stress) are also dealt with briefly. Concerning paralinguistic variation, ongoing work is presented on different voice qualities as produced by a single speaker. In conclusion, some implications for voice source rules in speech synthesis are mentioned.

1. INTRODUCTION

The work reported here addresses certain aspects of voice source variation in running speech. It aims to contribute to our understanding of the voice source and of its linguistic and paralinguistic functions. These two are interrelated: whereas intentional long term voice quality changes may be used to signal state of mind and attitude, even in modal voice there may be considerable modulation of quality as a function of linguistic factors. A further motivation for this work is the hope that this kind of information may have useful applications in speech technology. Specifically, one must understand the nature and function of voice source modulations if one is to capitalise on the more sophisticated source models currently being experimented with in speech synthesis.

Below are described some results from past and ongoing work by the two authors: Section 3 deals with source co-articulation to adjacent voiceless/voiced consonants; Section 4 discusses certain consequences of differences in stress pattern; and Section 5 presents preliminary measures concerning voice quality variation.

2. METHODS

The main analysis technique involves inverse filtering of the speech waveform. In order to obtain quantifiable results, a parametric model of differentiated glottal flow (the LF-model, described in [1]) is matched to the output of the inverse filter. The four parameters estimated, which are dealt with in this paper, are EE, RA, RK and RG. The first two of these are illustrated in Fig. 1. EE is the excitation strength and is measured as the negative amplitude of the differentiated flow at the moment of maximum discontinuity (Fig. 1a). It corresponds to the overall intensity of the signal, so that an increase in EE amplifies all frequency components equally (Fig. 1b). RA is a measure of the return phase, which is the residual flow from excitation to complete closure. The acoustic consequence of the return phase is a steeper spectral slope. As can be seen in Fig. 1b, a large RA corresponds to greater attenuation of the higher frequencies. RK is a measure of the skew of the glottal pulse: a larger value means a more symmetrical pulse shape. RG is a measure of the opening branch of the glottal pulse. RK and RG together mainly determine the level of the lower harmonics in the source spectrum. For a more detailed exposition on source parameters, see [2].

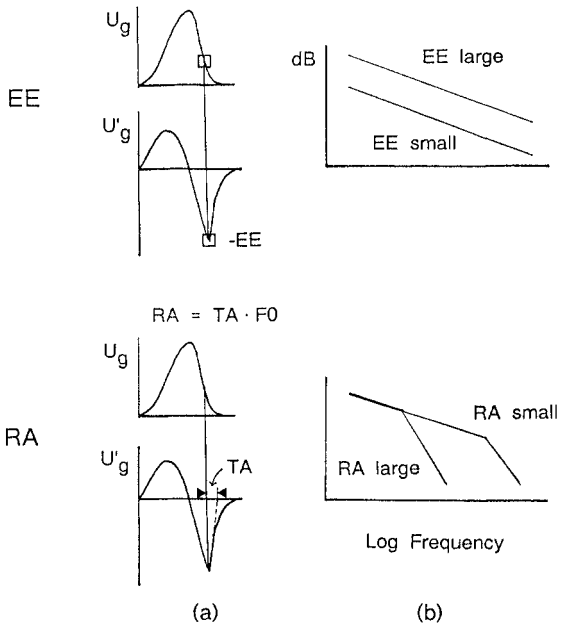

Fig. 1. For the source parameters EE and RA are schematically illustrated: (a) true and differentiated glottal flow, and (b) the effect on the spectrum of changing the value of either parameter.

As a complementary technique, the levels of F0 and the first four formants were measured from narrow-band spectra. In Section 3, information was needed on the relative timing of glottal and supraglottal gestures and on incomplete glottal closure during the vowel. For this, oral airflow was recorded using a Rothenberg mask. A more extensive description of all recording and analysis procedures used, can be found in [2].

3. COARTICULATORY EFFECTS OF VOICELESS/VOICED SEGMENTS

The voice source of the vowel may be affected by the voiced/voiceless nature of an adjacent consonant. The effects on the vowel of preceding or following stops or fricatives were studied in ¹CVCV nonsense utterances for Swedish, French, English, German and Italian (detailed results on the first three languages are presented in [2]). Stops in these languages include three main categories: voiceless aspirated, voiceless unaspirated and voiced.

The most striking cross language difference found concerned the offset of the vowel preceding a voiceless stop. Note for the Swedish data in Fig. 2 that the first vowel is greatly affected by whether the medial consonant is voiced or voiceless: the voiceless consonant causes weakening of EE early in the vowel. Concomitant with the drop in EE there is a rise in RA indicating a more rounded closing section of the glottal pulse as the vowel progresses. Measurements of spectral levels show extensive attenuation of formant levels, whereas the level of F0 remains relatively constant. Effectively, in the course of the vowel the mode of phonation becomes increasingly breathy voiced. A short interval of voiceless aspiration is frequently also observed before oral closure. These offset characteristics are essentially the same as for those stops traditionally described as preaspirated, as for example in Icelandic or Scottish Gaelic [3]. In contrast to Swedish, results for French show that the voiced/voiceless nature of the medial consonant has little effect on EE and RA values of the preceding vowel (Fig. 2). Spectral levels are similarly unaffected.

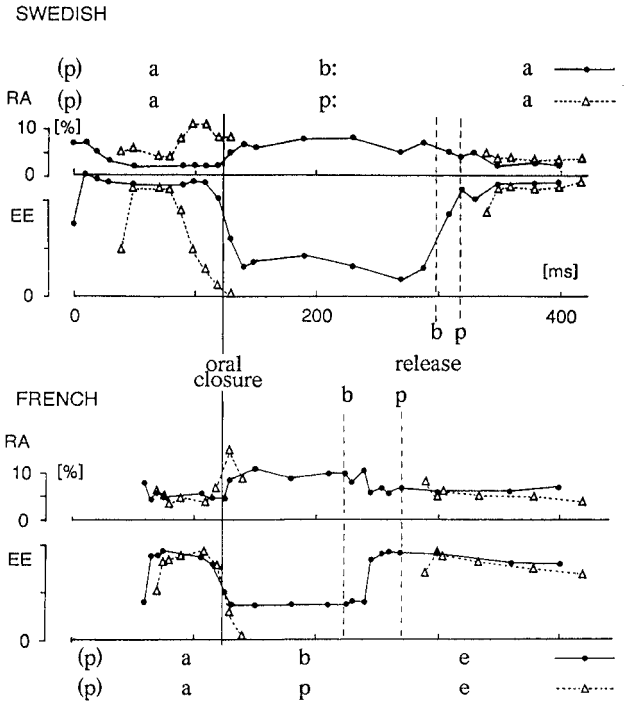

Fig. 2. In Swedish (above) and French (below) EE and RA values are superimposed for sequences where the medial stop is either voiced (solid lines) or voiceless (dotted lines). Oral closure is shown by a vertical solid line, and release by vertical dashed lines.

These source differences for French and Swedish reflect differences in the coordination of oral and laryngeal gestures. Oral airflow recordings for Swedish show a sharp rise in airflow rate during the vowel. From this we infer that laryngeal abduction begins very early relative to the oral occlusion (which can be identified from the sharp drop in oral airflow). In French, oral and laryngeal gestures would appear to be more tightly synchronised; oral airflow rate during the vowel is little affected by the nature of the following consonant.

The nature of a preceding consonant had relatively little effect on the onset characteristics of the vowel in most of the data studied. In Fig. 2, one can observe for both languages that the source parameter values are very similar at the onset of the second vowel regardless of the nature of the preceding (medial) stop. The most typical onset patter for EE was a sharp rise, no matter whether it followed a voiced, voiceless aspirated or voiceless unaspirated stop. For aspirated stops, however, some attenuation in the F1 region is observed for some time, and RA can be slightly higher. These observations, together with the fact that the oral airflow rate is still high for the first few glottal pulses following the aspirated stop, suggest that the glottis is not maximally adducting at onset. Yet, note that the effect is relatively minor compared to the slow offset found for Swedish before a voiceless stop (Fig. 2).

The German data was, however, notably different in this respect. Following the aspirated stop, the typical onset pattern was a gradual rise from a low EE and a gradual lowering from a high RA. This slow breathy voiced onset does seem very similar to the slow offsets typical for Swedish. Sporadic instances of similarly slow onsets were also found following aspirated stops in the other languages.

The contextual effects of a voiceless consonant can be very different depending on whether it precedes or follows the vowel. If the abduction/adduction gesture for a voiceless stop occurs while the vocal tract is unoccluded, offset/onset characteristics may be rather different. The offset of voice would appear to necessarily involve a gradual transition with breathy voiced phonation. The onset may involve such a similar gradual transition (as in German), but is much more rapid and efficient in most of the cases in the present study. This is clearly something the speaker has control over, though languages may tend to prefer one or other strategy. (Note that these points do not cover the case of glottalised voiceless stops.)

4. SOME PROSODIC EFFECTS

Source characteristics may vary as a consequence of the prosodic pattern. Some data on the effects of varying the stress pattern of a sentence are reported in [4], based on repetitions (by three speakers) of the Swedish utterance *Vi vill behålla honom*, elicited so that emphatic stress alternately fell on *vill, behålla* and *honom*. The word *behålla* /bɛˈhɔlːa/ could thus be analysed for postfocal, focal and prefocal environments.

The parameter that exhibited the most consistent variation with stress context was EE. For all speakers, EE exhibited a larger dynamic range in focal position: vowels are typically stronger and consonants are typically weaker. Most of the increase in dynamic range is due to the lowering of EE for the consonants /h/ and /l/. The very weak EE for /h/ corroborates electroglottographic results presented in [5], which suggest that the degree of glottal abduction for the voiceless consonant is positively correlated with the degree of stress.

It seems, furthermore, to be the case that an unstressed vowel is more affected by a following (emphatically) stressed element than a preceding one. The EE level of the vowel immediately preceding the emphatically stressed syllable is

also raised, whereas for the vowel in a following unstressed syllable it tends to be relatively lowered.

5. VOICE QUALITY VARIATION

Ongoing work also includes an attempt to describe the correlates of voice quality variation within the speech of a single speaker. The voice qualities we have investigated correspond to the descriptions by Laver [6], and include the following (note non-pathological) qualities: modal (neutral), tense, lax, breathy, whispery and creaky voice. The data on which work is being carried out comprises repetitions of a prose passage and of nonsense words inserted into a sentence frame, using the voice qualities just listed. The informant was a trained phonetician who is a native speaker of English and well versed in the Laver classification. For detailed description of the physiological correlates of the different voice qualities, see [6].

An initial approach to the task was to provide templates of some of the main voice qualities. This was done by comparing for different voice qualities the initial vowel of the nonsense word /ˈbæbə/ pronounced with nuclear stress. Some of these results are described in [7].

One problem with this approach is that a switch between voice qualities may not involve a single transformation, which remains uniform throughout an utterance. The differences between voice qualities may be much more pronounced in some environments than others. For example, the effects of a breathy voice quality may be less in evidence in the strongly stressed (nuclear) environment than in the unstressed one. Similarly, for creaky voice, creak does not appear throughout, but appears to be associated with specific environments, such as unstressed syllables, phrase terminations, voiced segments which have considerable supraglottal occlusion, etc. Therefore, as a refinement on the first approach mentioned, analysis is currently being carried out on a wider range of contexts. The observations below and Figs. 3, 4 and 5 are based on detailed analysis for all six voice qualities of the unstressed word /straiks/ taken from the prose passage.

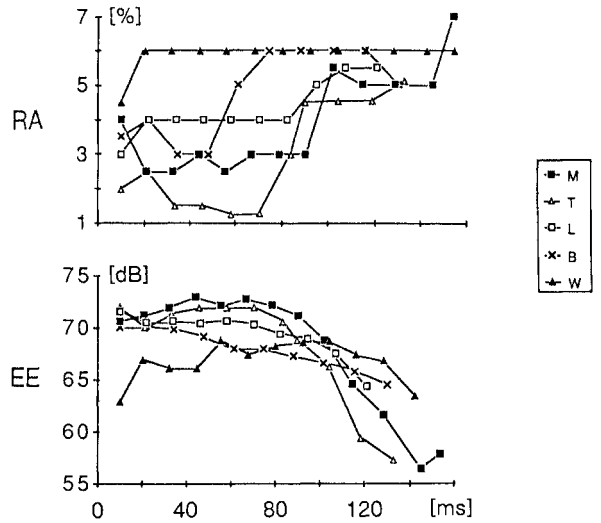

Fig. 3. RA and EE values for modal (M), tense (T), lax (L), breathy (B), and whispery (W) voice qualities in unstressed /straiks/.

Fig. 3 shows RA and EE values in the voiced portion of the word, for all qualities excepting creaky voice. Fig. 4 shows the relationship of F1 and F2 levels to that of F0 in tense, modal, lax and breathy voice for an interval of 90 ms. Fig. 5 illustrates RA and EE values for creaky (C) and modal (M) voice. The speech waveform is also shown for creaky voice.

Tense voice has, perhaps surprisingly, a lower F0 than has modal voice. The relatively very low RA indicates that this voice quality has the shortest return phase (and therefore the flattest spectral slope) of the qualities studied. EE is rather high, but note not quite as high as for modal voice (which has a higher RA). This is also somewhat surprising as one generally expects EE to be negatively correlated with RA. As can be seen in Fig. 4, F1 and F2 dominate F0. This effect may be partially explained by the RA parameter value, but almost certainly also reflects the very narrow bandwidths which characterise this voice quality.

Lax voice exhibits RA values considerably higher than modal, showing that the spectral slope is greater. The EE values, however, are rather high, very close to those found for tense voice. F0 is more prominent (than for tense or modal voice), being roughly at the same level as F1 (see Fig. 4). This again reflects the effect of the RA parameter and bandwidths.

Breathy voice is characterised by high RA values, rather low EE and low F0. The high RA, indicating dynamic leakage, would of course be expected in this mode of phonation. Bandwidths are wide, particularly for the lower formants. In Fig. 4, note that the spectrum is now clearly dominated by F0 and there is sharp attenuation of frequencies above this region. The relative dominance of F0 (and lower harmonics) in the spectrum is reflected also in the low RG and high RK values (and thus a very high open quotient) obtained for this voice quality.

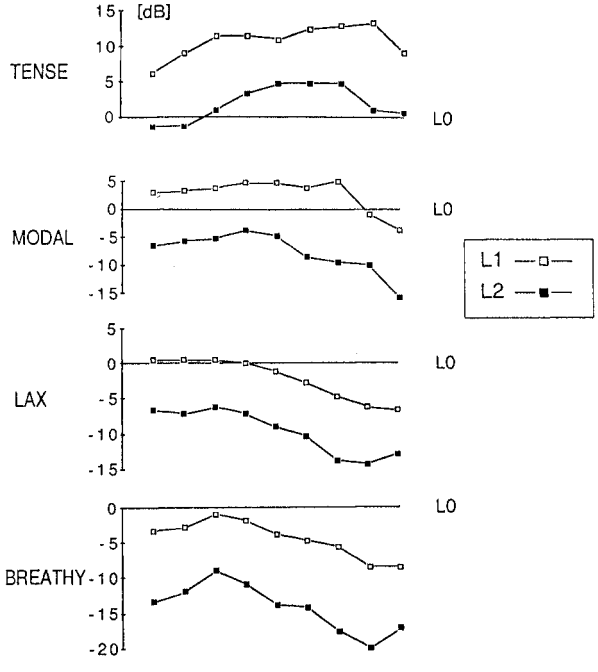

Fig. 4. F1 and F2 levels (L1 & L2) shown in relation to that of F0 (L0) for tense, modal, lax and breathy voice.

Whispery voice shows the most extreme RA values. As with breathy voice F0 dominates the spectrum, though its level is slightly less. RG is fairly similar to the breathy voice value,

but RK is considerably lower. The open quotient, though high, is not as large as for breathy voice, a fact which can be attributed to the relatively low RK.

Creaky voice exhibits a striking alternation of short and long periods (see Fig. 5). Correlated with this, RA and EE show similar fluctuations: very low RA values are associated with the higher EE values. The pulse-by-pulse fluctuations in RA cover the entire range found in the other voice qualities.

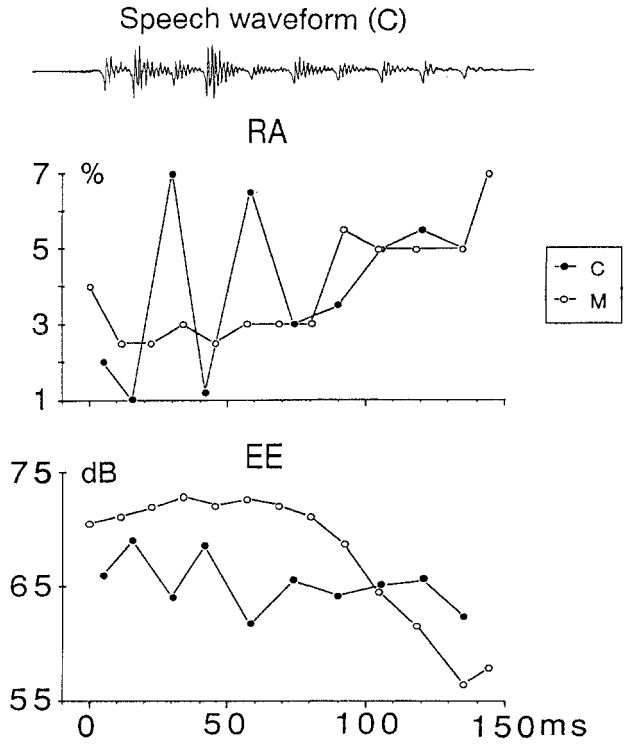

Fig. 5. RA and EE values for creaky (C) and modal (M) voice in /straiks/. Speech waveform also shown for (C).

Note, however, that these characteristics for creaky voice were specific to this environment. In the stressed environment examined, this alternating pattern of pulses was not observed and there was relatively little difference in the measured parameters for creaky and modal voice. Breathy voice values were also much closer to modal values in the stressed syllable.

As can be seen in Figs. 3 and 5, source values are not constant throughout the voiced sequence. RA values tend to converge towards a high value in the last pulses preceding the voiceless stop /k/. Concomitant with the rise in RA, EE values drop. This would appear to be a context effect such as is described in Section 3.

6. CONCLUSIONS: IMPLICATIONS FOR SYNTHESIS

One aspiration for this kind of descriptive work is that results may contribute towards a more natural source in synthesized speech. In recent years sophisticated source models have been devised, e.g. [1], to replace the simplistic impulse train hitherto used. To maximise on the benefits on these improved models we need extensive knowledge on the nature and function of the how true source varies in natural running speech. Work is currently ongoing in KTH to incorporate results such as these into a text-to-speech system. Space would not permit elaboration on detailed rules, so we will limit ourselves here to a few comments on the broad structure of the source rules component of a text-to-speech system.

• Linguistic rules would at least include the following: *Segmental rules* (not dealt with in this paper) should reflect the fact that the various supraglottal configurations for different vowels and for different voiced consonants affect aspects of the glottal pulse. Such rules might be expected to be universal as they would ultimately derive from production constraints. *Contextual rules* should capture the coarticulatory effects between segments. The content of Section 3 illustrates such contextual effects. It further exemplifies that contextual rules may at times need to be language specific. *Prosodic rules* will interact with both segmental and contextual rules. As illustrated in Section 4, the intrinsic source parameters of consonant and vowel segments will vary as a consequence of the stress pattern of an utterance.

• Paralinguistic voice quality rules. It has often been assumed that a change in voice quality should be implementable in terms of a single set of transformations from modal voice (which is of course modulated according to the linguistic rules). These transformations would remain uniform throughout the utterance. Though the work presented in Section 5 is preliminary, we would suggest that this is not likely to be the case. To effect a voice quality switch in synthesis, one would probably also need context sensitive rules to reflect the fact that the distance from modal voice will vary depending on the environment.

Acknowledgement

This work was supported by ACCOR, ESPRIT II, BRA no. 3279 and by a Swedish Institute scholarship.

References

[1] Fant, G., Liljencrants, J., & Lin, Q. (1985): "A four-parameter model of glottal flow", Speech Transmissions Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 4/1985, pp. 1-13.

[2] Gobl, C. & Ní Chasaide, A. (1988): "The effects of adjacent voice/voiceless consonants on the vowel voice source: a cross language study", Speech Transmissions Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR/2-3 1988, pp. 23-59.

[3] Ní Chasaide, A. (1985): "Preaspiration in phonological stop contrasts", unpublished Ph.D. thesis, University College of North Wales, Bangor, March 1985.

[4] Gobl, C. (1988): "Voice source dynamics in connected speech", Speech Transmissions Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 1/1988, pp. 123-159.

[5] Ní Chasaide, A. (1987): "Glottal control of aspiration and of voicelessness", Proc. 11th Int. Cong. of Phonetic Sciences, Tallinn, Vol.6, pp. 28-31.

[6] Laver, J. (1980): *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge.

[7] Gobl, C. (1989): "A preliminary study of acoustic voice quality correlates", Speech Transmissions Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 4/1989, pp. 9-22.