951

In O. Fujimura & M. Hirano (Eds.) Vocal Fold Physiology: Voice Quality Control.
Singular Publishing, Inc.: San Diego. (pp.3-22).
(1995)

CHAPTER
1

Voice Source Variations in Running Speech: A Study of Mandarin Chinese Tones

Anders Löfqvist, Laura L. Koenig, and Richard S. McGowan

1. INTRODUCTION

This paper examines variations in the sound source during naturally produced utterances by inverse filtering of the oral/nasal airflow. The relations between stress, tone, and intonation and the voice source have primarily been examined using acoustic analysis of fundamental frequency, and there exists a large body of data on F0 variations in speech. Aerodynamic studies of voice source in relation to changes in intensity and fundamental frequency have mostly been concerned with steady phonations. Early studies of airflow were restricted to average flow rates because of the limited bandwidth of the recording systems. Recording systems for airflow with a higher frequency response, such as the one described by Rothenberg (1973), in combination with inverse filtering procedures using digital signal processing techniques (cf. Javkin, Antonanzas-Barroso, & Maddieson, 1987) have recently allowed more detailed analysis of voice source aerodynamics. For example, Holmberg, Hillman, and Perkell (1988, 1989) examined source pulse properties as a function of intensity and fundamental frequency. Such controlled studies are necessary for clarifying the basic relationships among the source aerodynamic parameters. However, because fundamental frequency and intensity often covary in speech, it is also important to examine these relationships in naturally produced utterances.

A few studies have examined source properties in running speech using inverse filtering of the acoustic signal. Results presented by Gobl (1988) suggest that the excitation strength of the source pulse (taken as the magnitude of the derivative of glottal flow during the closing phase) is higher in stressed than in unstressed vowels. Pierrehumbert (1989) reports that the excitation strength varies with the fundamental frequency pattern of the utterance as well as with the overall voice level. Pierrehumbert and Talkin (1992) show that the ratio of the fundamental to higher harmonics is affected by accent and by position relative to an intonation phrase boundary.

The present experiment was designed to provide further information on voice source properties in naturally produced utterances. To obtain records of different F0 patterns in a controlled and systematic fashion, the tones of Mandarin Chinese were used as speech material. Mandarin is traditionally analyzed as having four lexical tones (cf. Howie, 1976; Tseng 1981; Xu, 1993). Thus, a syllable can have up to four different meanings, depending on the tone used with it. Because fundamental frequency contours serve to make linguistic distinctions in Mandarin, we reasoned that F0 variation in an utterance would be constrained, and that this would provide a way of eliciting controlled differences in pitch contour. The traditional descriptions of the Mandarin tones are as follows: Tone 1 is described as high-level, tone 2 as mid-rising, tone 3 as low-rising, and tone 4 as high-falling. Figure 1–1 shows the acoustic signal and fundamental frequency curves for single productions of each of the four tones, as produced by our female subject FW.

These voice source data are interpreted in terms of the cover-body models of the vocal folds (Hirano, 1981; Fujimura, 1981; Titze, Luschei, & Hirano, 1989; Titze, 1991; Farley, 1994). In a cover-body model, for a fixed subglottal pressure and state of glottal adduction, the factors that determine fundamental frequency—length and tension of the cover of the folds—are themselves determined by the activities of the cricothyroid muscle and the thyroarytenoid muscle. Muscle activation plots (e.g., Titze, 1991) show that the fundamental frequency does not uniquely specify the activities of these muscles. However, because muscle activations help to determine voice quality, as well as fundamental frequency, measures such as the open quotient (OQ) may help clarify

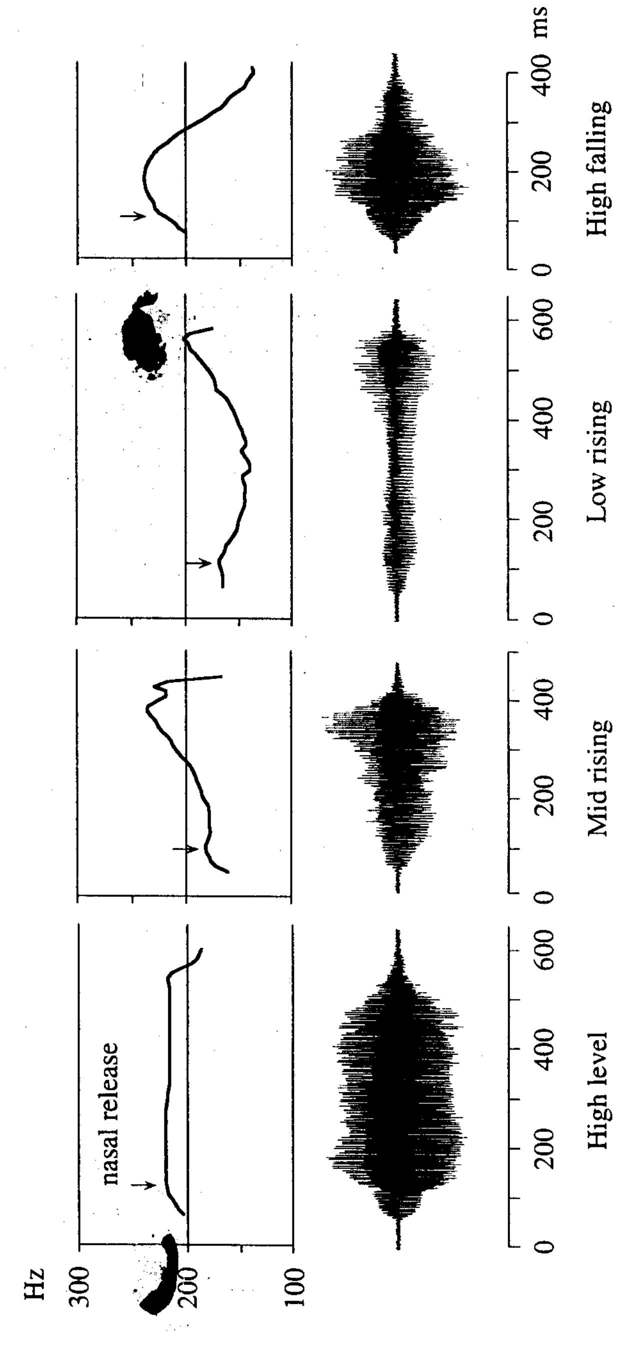

Figure 1–1. F0 and acoustic signal for single productions of Mandarin tones uttered by speaker FW.

the levels of muscle activation for a given fundamental frequency. Of course, source characteristics may not be determined solely by these intrinsic muscles of the larynx. Other influences may include activity of the sternothyroid (Niimi, Horiguchi, & Kobayashi, 1991), sternohyoid (Sawashima & Hirose, 1983), and cricopharyngeus (Honda & Fujimura, 1991); the subglottal pressure; and the state of vocal fold adduction. The changes in some of these variables can be inferred from measures of pulse amplitude and time average flow. Thus, we generally should be able to infer more about laryngeal muscle activity with more measures of the glottal volume velocity than just fundamental frequency.

2. PROCEDURE

2.1. Data Recording

Airflow was recorded using a face mask and a differential pressure transducer (Glottal Enterprises) according to the method described by Rothenberg (1973). The system has a flat frequency response from DC to above 1 kHz (cf. Badin, Hertegård, & Karlsson, 1990). The flow signal for each subject was calibrated using a rotameter. During the experiment, the flow and audio signals were digitized at 10 kHz with 12-bit resolution. Because the mask covers the mouth and the nose, the flow signal represents both oral and nasal airflow.

Two native speakers of Chinese, one male (WN) and one female (FW), produced the four tones of Mandarin Chinese on the single syllable "ma". Each of the four combinations is an existing word in Mandarin, so that the stimulus set comprised a four-way minimal contrast. Each word was repeated up to 20 times in isolation, with a short pause between each repetition.

2.1.1. Data Processing

After sampling, portions of the flow signal were extracted for software filtering. Linear predictive coding with covariance analysis and root finding was used to estimate the parameters for removing ripple in the flow pulses. Before each file was filtered, the open and closed phases of each pulse were marked by hand from the flow file.

Default values for the analyses were used except where the filtering resulted in spurious high frequency formants or no first or second formants. The default order of the analysis was 12 and the window size was 80 points (8 msec). A window of this size contains both a maximum

and a minimum of the glottal pulses. In cases where the analysis did not work (e.g., the resulting pulses were not smooth or the program crashed), the order of analysis could be reset to 10, or the window size could be reset to anywhere from 60 to 120 points. Root-solving was used to find the formant frequencies and bandwidths to produce the inverse filter. Various parameters were set to exclude spurious formant frequencies at this stage. These included the minimum and maximum frequencies of 500 and 5000 Hz, respectively, as well as the minimum and maximum bandwidths of 50 and 500 Hz. Although 50 Hz is a reasonable bandwidth for a first or second formant in a closed glottis condition (Fant, 1961), the analysis window was long enough so that the first formant was both damped and moved by subglottal coupling, both effects adding to the apparent bandwidth (Fant, 1972). Setting the minimum bandwidth lower would produce poles at harmonics of the voice.

Each analysis window used to compute the filter started at some time during the closed phase. A new filter was computed every 15 samples, as long as the earliest point in the analysis window was in the closed phase. By requiring the window to start in the closed phase, we tried to minimize effects due to changing relative phase between window and voice pulse.

2.1.2. Measurements

In preparing the filtered flow signal for measurements, relevant points in the waveform were labeled interactively on a computer. As illustrated in Figure 1–2, these points were the onset (T1 and T4) and offset (T3) of the open phase during each period and peak flow (T2). Points T1 and T3 were defined on the basis of a rapid change in the signal. This definition of the closed phase allows for flow occurring during the "closed" phase. Based on these labels, the following measurements were extracted for a predetermined number of periods. The value of peak flow was measured at point T2. Minimum flow for each cycle was taken as the average of the flow at points T3 and T4. Pulse amplitude was taken as the difference between maximum and minimum flow. The open quotient was calculated as the ratio between the open phase, T1–T3, and the period time, T1–T4. The area of the pulse was obtained in the following way. First, the flow signal was integrated between points T1 and T3. Second, a rectangle was defined with its length equal to the duration of the open phase (T1–T3) and its height equal to the mean of the flow at points T1 and T3. Third, the area of the rectangle was subtracted from the value obtained by integration. Thus, the area of the pulse represents the area of the open phase with the base, or DC, flow removed.

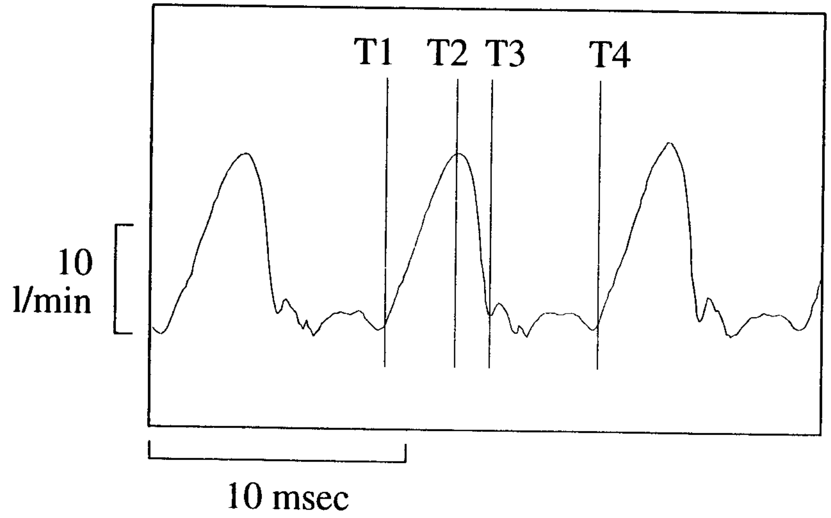

Figure 1–2. Measurement points on filtered flow.

Because the number of pitch periods varied as a function of speaker, utterance, and token, and also given processing limitations, the onsets and offsets of the vowels were selected for analysis. The transition between the initial nasal and the vowel was determined according to a change in formant structure and a shift in overall amplitude. The end of the vowel was taken as the end of periodic pulsing. We assumed continuity and monotonicity in F0 and source properties in the middle regions of the vowels, between the onset and offset regions. Note in the subsequent figures that the last onset measurements are generally quite similar to the first offset measurements. For tones 1, 2, and 3, 20 pitch periods were measured after vowel onset, and 30 periods were measured preceding vowel offset. (More periods were measured at the offset because some utterances showed fairly long regions of low-amplitude pulsing at the end, and we wanted to be sure to measure back into the true vocalic region.) For both speakers, the vowel duration for tone 4 was much shorter than durations for the other tones. For FW, the 50 measured periods constituted virtually all of the vowel. With WN's lower fundamental frequency, some vowels for tone 4 contained no more than 30 periods. Here, taking the shortest token as the limiting case, the first 15 pulses were taken as onset measurements, and the last 14 were taken as offset.

Once measurements had been made for 12 tokens, average values were calculated on a pulse-by-pulse basis (averages were obtained separately for the first, the second, the third pulse and so on). Pulses preceding the offset of the vowel were handled in a similar manner, except that separate averages were obtained for the last pulse, the last pulse minus one, and so on. Time scales for each utterance were obtained as cumulative averages. Averaging serves to smooth out small variations caused by measurement error and pulse shape variations. Comparisons of the averaged plots with token plots verified that the averages fairly represented the trends for individual tokens, except in cases where rapid change in source properties yielded short-term fluctuations in temporal measurements for individual tokens.

3. RESULTS

The results for FW and WN are shown in Figures 1–3 through 1–8. Figures 1–3 and 1–4 show representative tokens of unfiltered flow for each tone for each speaker. The fundamental frequency, open quotient, minimum and maximum flow, and area and amplitude of glottal pulse are given as time series of averages over 12 tokens of the filtered flow signals (Figures 1–5 through 1–8). The axes for WN possess different ranges from those of FW because WN's flow rates were generally higher and his fundamental frequencies were generally lower than FW's.

The shapes of the fundamental frequency contours are qualitatively similar for both subjects, and the patterns of F0 change are consistent with previously published sources, such as Howie (1976). The names of the tones are fairly descriptive of these contours. Note that the low-rising tone shows a falling F0 during vowel onset; this tone is sometimes called falling-rising. For both of the subjects, the voiced portion of the high-falling tone (tone 4) was of shorter duration than the other tones. The mid-rising tone also tended to be relatively short, especially for subject FW.

In the analysis that follows, the effects of subglottal pressure variations and external laryngeal muscles are ignored. It will be shown that the observed tonal variations might be accounted for in terms of abductory-adductory activity and in terms of thyroarytenoid (TA) activity and cricothyroid (CT) activity alone. This, of course, provides no proof that such is indeed the case, but it offers a plausible scenario. In this account, a smaller open quotient could be caused by greater adduction due to increased activity of the TA and/or other adductory muscles. Alternatively, it could be caused by a laxer cover resulting from decreased

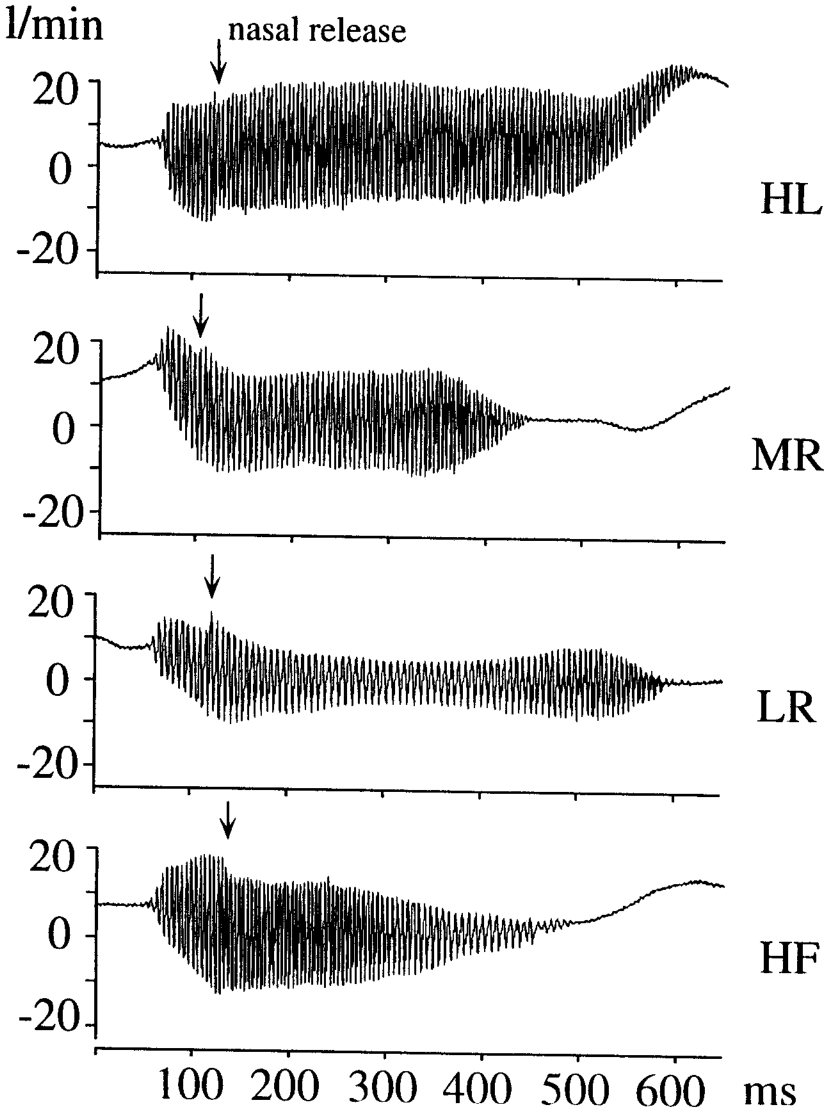

Figure 1–3. Unfiltered flow signals for single tokens of the four tones uttered by speaker FW.

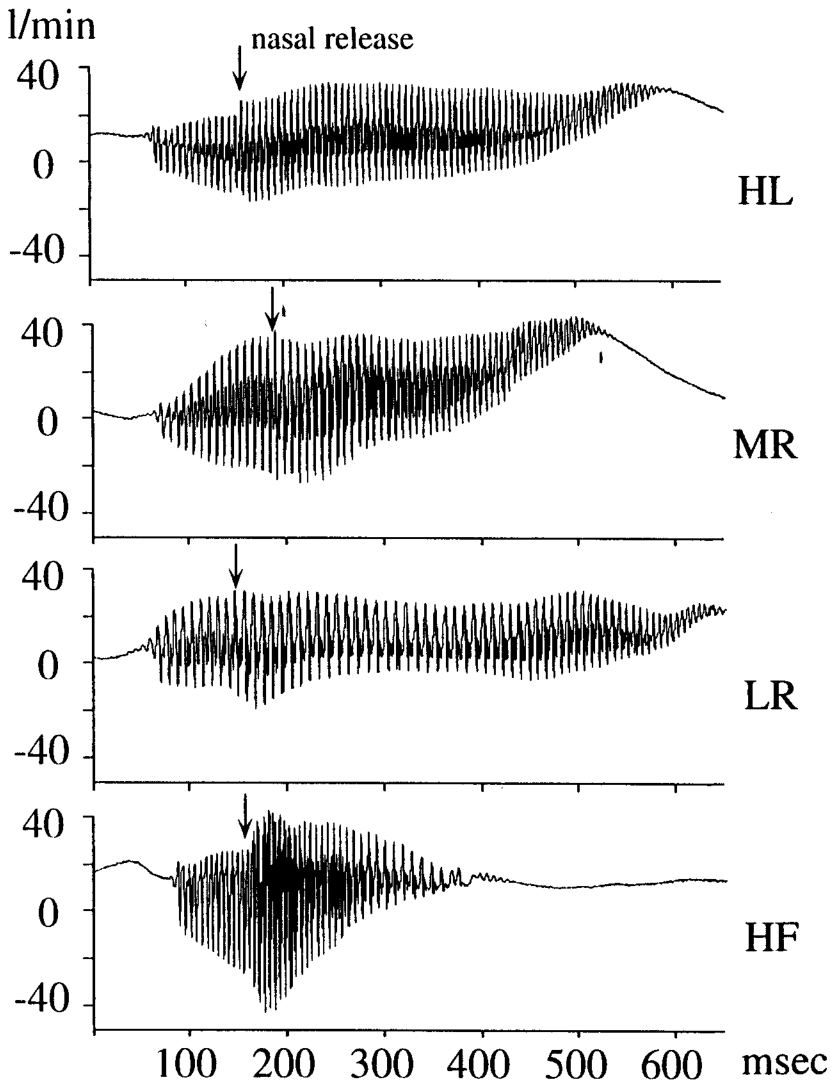

Figure 1–4. Unfiltered flow signals for single tokens of the four tones as uttered by speaker WN.

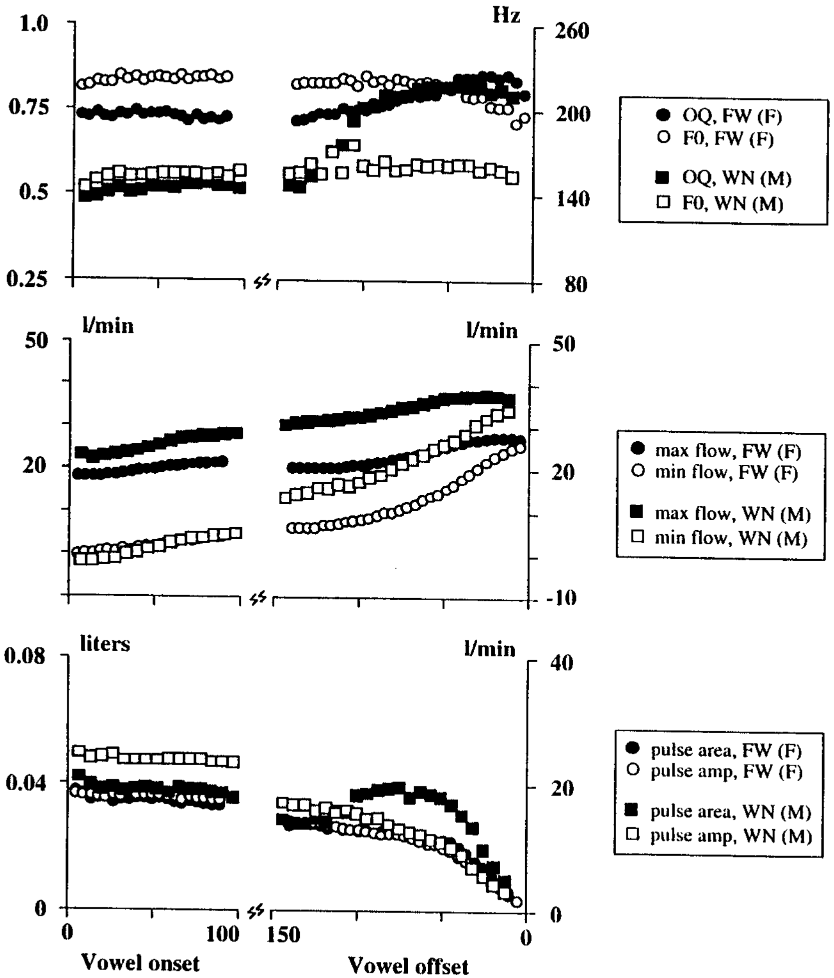

Figure 1–5. Time series of measurements for tone 1 (high-level).

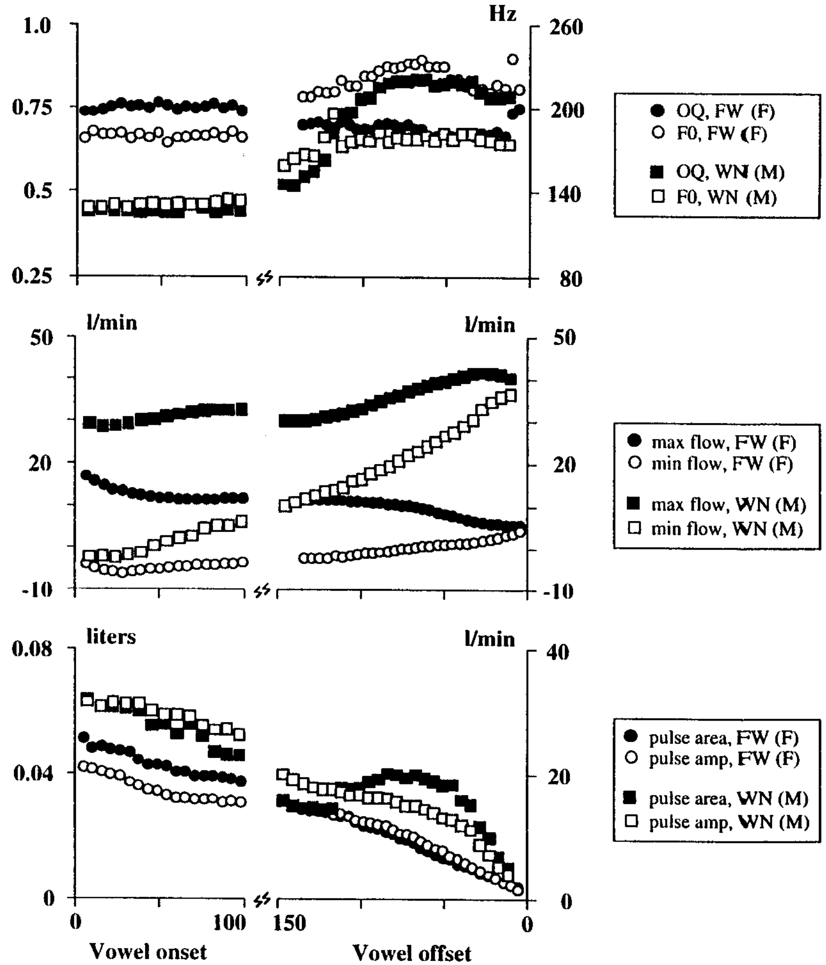

Figure 1–6. Time series of measurements for tone 2 (mid-rising).

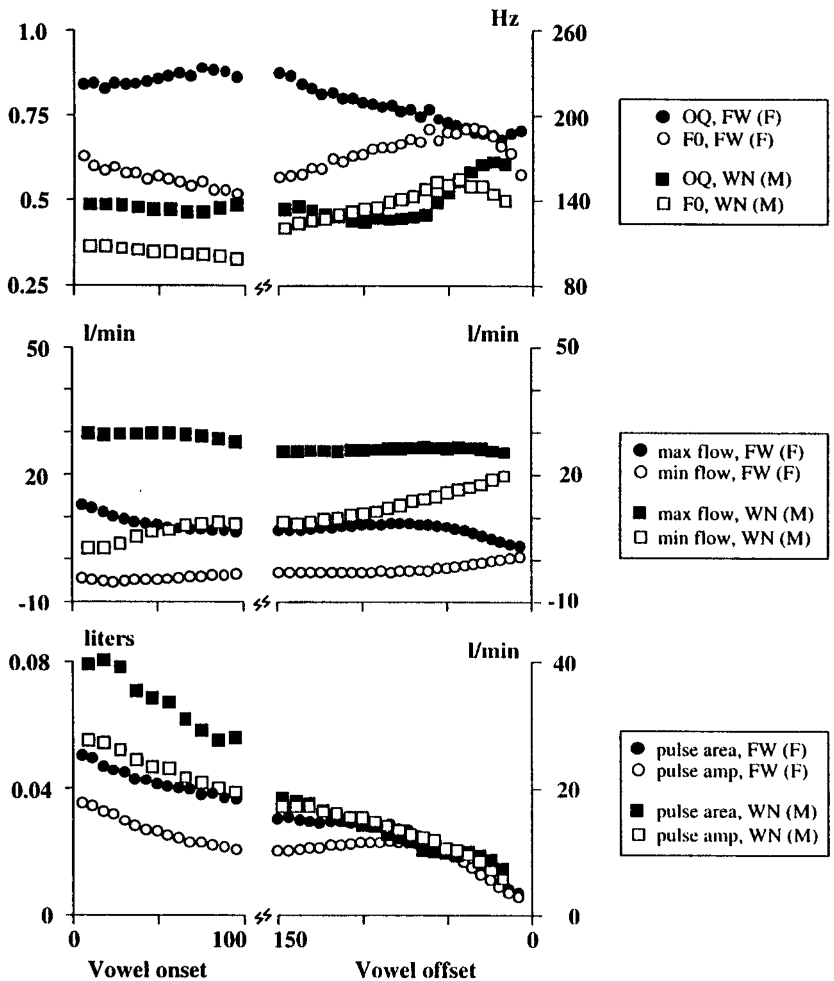

Figure 1-7. Time series of measurements for tone 3 (low-rising).

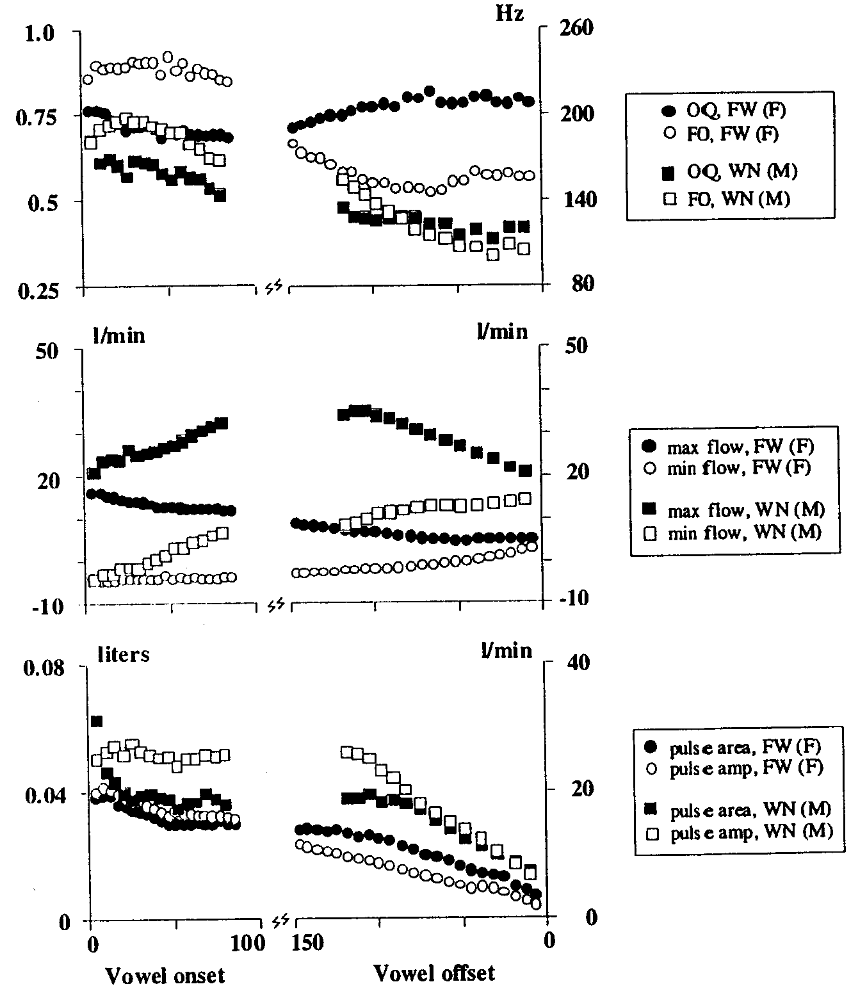

Figure 1-8. Time series of measurements for tone 4 (high-falling).

CT activity and/or increased TA activity. Increases in minimum flow and open quotient with an increasing or steady maximum flow, and decreases in pulse amplitude indicate abduction of the vocal folds.

The trends for the high-level tone (tone 1) are similar for both WN and FW. Indeed, the fundamental frequency is relatively steady for both subjects (Figure 1–5). However, they both exhibit large increases in open quotient toward the end of the utterance. They also show steady increases in minimum and maximum flow through the utterance, and decreases in pulse amplitude. These changes are attributed to abduction of the folds through the utterance, and the representative tokens in Figures 1–3 and 1–4 show such endings. In terms of the cover-body picture of the folds, there seems to be little adjustment in the CT and TA necessary, except, perhaps, to counteract any lengthening of the folds caused by abduction.

For the mid-rising tone (tone 2), WN and FW appear to have obtained similar frequency contours using two different laryngeal strategies. Again, WN shows signs of vocal fold abduction through the utterance, with increasing open quotient, minimum and maximum flows, and decreasing pulse amplitude (Figure 1–6). The increase in fundamental frequency and open quotient is consistent with increasing CT activity and, perhaps, decreasing TA activity. FW, on the other hand, has a pattern of gradual decrease in maximum flow and open quotient, but with increase in minimum flow. Increased fundamental frequency with no increase in open quotient could arise from increasing CT and TA activity with, perhaps, a slight adduction. Note that the combination of TA activity and activity from other adductory muscles would lead to a decrease in open quotient. With diminishing amplitude of vibration, less and less of the stiffness of the TA influences the vibrational frequency (Titze et al., 1989), so that TA cannot be activated too much because that would counteract the effect of the CT in raising fundamental frequency.

For the initial part of the low-rising tone (tone 3), both FW and WN show declines in fundamental frequency that might be attributed to diminishing CT activity (Figure 1–7). Whereas WN's open quotient stays constant, FW's increases. Decreased TA activity for less bunching of the cover could cause this trend for FW. From the mid part of the vowel, the overall trend for FW is like that of the mid-rising tone (tone 2), although here the open quotient starts higher, and fundamental frequency and maximum flow start lower at mid-vowel (Figures 1–6 and 1–7). Both CT and TA activity increase to achieve higher fundamental frequency and lower open quotient into the end of the vowel. Before the end of the vowel there is a slight rise in pulse amplitude, so that more of the body is involved with vibration and tensing the TA should increase the vibration frequency of the folds. Note there is a small decline in fundamental frequency at the very end of the vowel, which coincides with a decrease in vibration amplitude, when less of the body is involved with the vibration. In contrast to FW, WN again seems to abduct the folds, although here it is delayed until the very end of the vowel, yielding late rises in open quotient and minimum flow. As with WN's tone 2, CT activity appears to be responsible for the rise in fundamental frequency in tone 3. However, before the abduction becomes apparent, there appears to be activity in the TA, which has the effect of preserving open quotient.

For the high-falling tone (tone 4), WN shows a monotonic fall (except at the very beginning) in fundamental frequency and in open quotient (Figure 1–8). The pattern here seems to be one of low initial TA activity and high CT activity because of the relatively high open quotient for this speaker. Subsequently, TA activity remains constant or increases along with a rapid decrease in CT activity, so that open quotient may decrease with fundamental frequency. FW, on the other hand, shows a nonmonotonic change in open quotient. Open quotient initially decreases with fundamental frequency, which is similar to WN, but in the middle of the utterance it rises and then levels as the vowel is completed. This mirrors the fundamental frequency, which initially falls and then slowly rises at the end of the vowel. Evidently, the activation of the TA remains constant as an initially high CT activation decreases into the middle of the vowel. From mid-vowel on, both CT and TA activity decrease to produce falling fundamental frequency with increasing open quotient.

4. DISCUSSION

The most striking result of these data was that the fundamental frequency contours for each tone were of similar shape for both subjects, but the voice quality changes were not. In some instances, one subject would abduct at the end of some tones (namely, WN for tones 2 and 3), while the other would not. Further, even when subjects did not appear to be abducting differentially, the change in open quotient in one subject could be opposite from that seen for the other, as was the case with tone 4. Thus, our subjects used different laryngeal strategies in attaining the same fundamental frequency contour.

Because this analysis only considered the cover-body picture of the larynx, some of the other mechanisms for fundamental frequency control were not considered. External muscle activations or movements of the larynx that change the relative positions of the thyroid and cricoid cartilages may also have effects on voice quality. Thus, what has been attributed to the cricothyroid here could be due to many other factors. Further, it remains to be verified that subglottal pressure remains constant, although given the short durations of these utterances there should not be much active alteration of this quantity during single productions of any utterance. Finally, the cover-body models as currently implemented rely on a string model of vibration, so that any beam properties, such as bending stiffness or shear modulus, are neglected (Johnson, 1992). It should also be mentioned that much of cover-body theory has been formulated on measurements made from canine larynges. Given the differences between canine and human larynges in structure and tissue characteristics, more speech data are needed to verify the specifics of the theory. Airflow is one noninvasive means of obtaining more information about laryngeal activity in humans. Two further steps in this research are to combine these data with electromyographic data on muscle activations and with results obtained from models of laryngeal mechanics. We expect such convergence of information to become indispensable in inferring laryngeal behavior.

ACKNOWLEDGMENT

This work was supported by Grant No. DC-00865 from the National Institute on Deafness and Other Communication Disorders.

DISCUSSION AFTER PRESENTATION

O. Fujimura: I was very impressed by the approaches taken in this paper, which touched upon the central concern of this voice quality conference: the change of voice quality within an utterance, within a sentence, even within a word. We have been neglecting this aspect of speech, except in some linguistic literature dealing with the phonemic use of register (phonation type) differences. This issue is crucial, not only for understanding Mandarin tones, but also many different language characteristics such as intonation, in addition to speaker characteristics, emotional state, and so on. I have many questions, and we will deal with some of the points I have in mind in the discussion session (see Section V).

J. Pierrehumbert: One comment and one question. The comment is that in Pierrehumbert and Talkin (1992), we had two speakers, and we did find that they treated the end of the utterance differently. One of

them was much more of an abductor and the other was much more of an adductor. So I'm thinking you might be seeing a boundary effect rather than the effect of the tones per se. It would be very interesting to get some data in a more continuous speech frame.

R. McGowan: I agree. They both seem to be adducting on tone 4 to stop the vibration.

J.P.: Let me say that the boundary interacts with the tone target you are trying to achieve, so it's not *only* the boundary.

The other thing is that I wondered if you could clarify the pertinence of Ingo's curves (Titze, 1991) to male and female voices.

R.M.: This is a good question. First, the physiological data in the curves are—stop me if I'm wrong, Ingo—based on canine data. In addition, we don't know the subglottal pressure and that changes the contours quite a bit, as he (Titze) shows. We wanted to get back to the laryngeal configuration using flow measures. At the end, as I mentioned, we don't have to attribute everything to CT and TA.

I. Titze: The curves should not be taken as being gospel in quantitative terms. They are generic curves that would apply to females, males, and dogs, for that matter.

J.P.: If you are in a certain F0, which part of the curve are you in? Are you down in the corner or out in the middle?

I.T.: That's the key. If you divide the plot into four quadrants, the main point is to show that there are different ratios between CT and TA activity that give you different slopes of the curves. That's the point we are making here.

J.P.: Physically you are presupposing that you are in the lower left quadrant for both of these people for the whole experiment.

R.M.: Right, for the whole experiment.

J. Sundberg: Do you have any idea of the perceptual relevance of these effects? I'm suspicious about what the implications should be in the case of singing, where you can do quite high quality synthesis of singing without these effects. I wonder how the singers are doing the equivalent of what you are finding, or maybe they don't sing in the same way in the languages you are investigating.

R.M.: I think you are asking the wrong person.

T. Ananthapadmanabha: Perhaps I can answer this question regarding the perceptual relevance of open quotient. I have conducted experiments on difference limen (DL) for OQ and SQ for steady state vowels.

I found the DL for OQ to be about 6% at 50% and about 10% at 70%. DL for SQ is about 40%. I will cover this aspect in my paper.

K. N. Stevens: These measures you are making are primarily low frequency measures. I wondered if you had looked at the spectra to see if there were any changes in return phase or tilt or whatever you want to call it? I know the inverse filtering won't give you this high frequency data, but presumably there would be some differences in the high frequency spectra for the different tones.

R.M.: No, we haven't. That's sort of a question somebody else asked, a question Osamu raised. If we had a higher frequency signal that we could trust above 1000 Hz, then we could get a speed quotient or something of that sort. Actually a speed quotient may be derivable here, but we don't have the maximum derivative, which should be important.

P. Ladefoged: I'm glad that you pointed out that speakers may differ in their mechanisms for controlling pitch. Some years ago I studied the mechanisms used for pitch change in Mandarin tones by (now) Professor Lin of Beijing Linguistic Institute. We recorded his subglottal pressure while producing a series of Mandarin utterances demonstrating two syllable combinations of tones. There was a very high correlation—almost perfect— between his F0 and his subglottal pressure. My memory is that his subglottal pressure varied over a range of about 7 cm H_2O, from about 6 cm H_2O to about 13 cm H_2O. Given that 1 cm H_2O may produce 4 Hz pitch change, this would correspond to a 28 Hz variation in his overall pitch, which was not much different from what was produced. I am sorry that I do not have the details of this unpublished experiment, as the data were left in Beijing. But it does seem to me that one must take variations in subglottal pressure into account, at the very least for some speakers of Mandarin.

REFERENCES

Badin, P., Hertegård, S., & Karlsson, I. (1990). Notes on the Rothenberg mask. *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report, 4*, 1–7.

Fant, G. (1961). The acoustics of speech. In L. Cremer (Ed.), *Proceedings of the Third International Congress on Acoustics, Stuttgart, 1959*. Amsterdam: Elsevier Publishing Company. Reprinted in G. Fant (1973), *Speech sounds and features* (pp. 3–16). Cambridge: MIT Press.

Fant, G. (1972). Speech analysis and speech production: Subglottal formants. *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report, 4*, 1–12.

Farley, G. R. (1994). A quantitative model of voice F0 control. *Journal of the Acoustical Society of America, 95*, 1017–1029.

Fujimura, O. (1981). Body-cover theory of the vocal fold and its phonetic implications. In K. N. Stevens & M. Hirano (Eds.), *Vocal fold physiology* (pp. 271–281). Tokyo: University of Tokyo Press.

Gobl, C. (1988). Voice source dynamics in connected speech. *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report, 4*, 123–159.

Hirano, M. (1981). *Clinical examination of voice*. Vienna: Springer-Verlag.

Holmberg, E., Hillman, R., & Perkell, J. (1988). Glottal airflow and transglottal air-pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America, 84*, 511–529.

Holmberg, E., Hillman, R., & Perkell, J. (1989). Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch. *Journal of Voice, 3*, 294–305.

Honda, K., & Fujimura, O. (1991). Intrinsic F0 and phrase-final F0 lowering: Phonological vs. biological explanations. In J. Gauffin & B. Hammarberg (Eds.), *Vocal fold physiology* (pp. 129–138). San Diego: Singular Publishing Group.

Howie, J. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press.

Javkin, H., Antonanzas-Barroso, N., & Maddieson, I. (1987). Digital inverse filtering for linguistic research. *Journal of Speech and Hearing Research, 30*, 122–129.

Johnson, M. A. (1992). Using beam elements to model the vocal fold length in breathy voicing. (Abstract). *Journal of the Acoustical Society of America, 91*, 2420–2421.

Niimi, S., Horiguchi, S., & Kobayashi, N. (1991). F0 raising role of the sternothyroid muscle: An electromyographic study of two tenors. In J. Gauffin & B. Hammarberg (Eds.), *Vocal fold physiology* (pp. 183–188). San Diego: Singular Publishing Group.

Pierrehumbert, J. (1989). A preliminary study of the consequences of intonation for the voice source. *KTH, Speech Transmission Laboratory, Quarterly Progress and Status Report, 4*, 23–36.

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Docherty & A. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 90–117). Cambridge: Cambridge University Press.

Rothenberg, N. (1973). A new inverse-filtering technique for deriving the glottal waveform during voicing. *Journal of the Acoustical Society of America, 53*, 1632–1645.

Sawashima, M., & Hirose, H. (1983). Laryngeal gestures in speech production. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 11–38). New York: Springer-Verlag.

Titze, I. R. (1991). Mechanisms underlying the control of fundamental frequency. In J. Gauffin & B. Hammarberg (Eds.), *Vocal fold physiology* (pp. 129–138). San Diego: Singular Publishing Group.

Titze, I. R., Luschei, E. S., & Hirano, M. (1989). Role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice, 3*, 213–224.