VOWEL AMPLITUDE VARIATION DURING SENTENCE PRODUCTION

Helen M. Hanson *

Sensimetrics Corp., 26 Landsdowne St., Cambridge, MA 02139, USA
M.I.T. Research Laboratory of Electronics, 50 Vassar St., Cambridge, MA 02139, USA

ABSTRACT

With the goal of synthesizing natural-sounding speech based on higher-level parameters, sources of vowel amplitude variation were studied for sentences having different prosodic patterns. Previous theoretical and experimental work has shown that sound pressure level (SPL) is proportional to subglottal pressure (P_s) on a log scale during production of sustained vowels. The current work is based on acoustic sound pressure signals and estimated P_s signals recorded during the production of reiterant speech, which is closer to natural speech production and includes prosodic effects. The results show individual, and perhaps gender, differences in the relationship between SPL and P_s, and in the degree of vowel amplitude contrast between full and reduced vowels. However, a general trend among speakers is to use subglottal pressure to control vowel amplitude at sentence level and main prominences, and to use adjustments of glottal configuration to control vowel amplitude variations for reduced and non-nuclear full vowels. These results have implications not only for articulatory speech synthesis, but also for automatic speech recognition systems.

1. INTRODUCTION

During the production of a sentence, there are significant changes in the amplitudes of the vowels and in the spectrum of the glottal source, depending on the degree of prominence that is placed on each syllable. Examples of the amplitude variations are given in Fig. 1, which shows a speech waveform for the utterance "Lisa's boy walked to school." The fourth syllable is emphasized, thus having a full vowel and phrasal stress, and it has the greatest vowel amplitude. The first and third syllables also have full vowels, but do not have phrasal stress, and their vowel amplitude is less than that of the fourth syllable. The second and fifth syllables have reduced vowels and even less amplitude. The sources of such amplitude variations (as well as corresponding spectrum variations) are of interest for the development of models of speech production, for articulatory speech synthesis, and for the design of methods for accounting for variability in speech recognition systems.

In particular, we are developing rules for HLsyn, a speech synthesis system in which the Klatt formant synthesizer [1] is controlled by a small set of higher level articulatory and formant parameters [2]. Although the quality of individual words is usually quite good, the synthesized speech is often unsatisfactory for utterances of more than a few syllables, because vowel amplitude does not vary in a natural way from one word to another. Therefore, we have an interest in developing synthesis rules that result in appropriate variation of vowel amplitude throughout an utterance. Currently, two HLsyn parameters affect vowel amplitude: subglottal pressure, P_s, and average glottal area, A_g, the latter controlling source spectral tilt, open quotient, and first-formant bandwidth. At this time, P_s is held constant throughout an utterance (although its value can be changed for a given utterance), while A_g, and thus the parameters it controls, can vary. Our goal, then, is to determine how these two parameters should be controlled in order to obtain natural vowel amplitude variation in an utterance.

Through a study of both the acoustic sound pressure signal and estimates of subglottal pressure from oral pressure signals, we have examined some of the sources of variation in vowel amplitude and spectrum during sentence production. In the next section we briefly review possible sources of vowel amplitude variation and some previous experimental results. Following that, we describe a speech production experiment in which we studied P_s and glottal configuration as sources of vowel amplitude variation in reiterant speech. Although previous work has concentrated on subglottal pressure as the source of vowel amplitude variation, our results suggest that glottal configuration also plays an important role in varying the amplitude between the three types of vowel described above. Some tentative rules for controlling vowel amplitude variation in synthesized utterances are suggested in the conclusion.

2. BACKGROUND

There are several possible sources of vowel amplitude variation. It has been shown theoretically [3] and experimentally (e.g., [4]) that sound pressure level (SPL) is proportional to $20 \log_{10} P_s^{3/2}$ for speech. The experimental studies have mainly relied on sustained vowels or repeated syllables. However, such utterances lack the natural variations in fundamental frequency, P_s, and glottal configuration that occur in normal sentence production. Another source of vowel amplitude variation is changes in the fundamental frequency (F0). If only F0 varies, then $SPL \sim 20 \log_{10} F0^{1/2}$. A third source is variation in vocal-tract losses: if the first-

*This work is supported by research grant numbers DC00075, MH52358, and 1 F32 DC 00205-02 from the National Institute on Deafness and Other Disorders, National Institutes of Health.

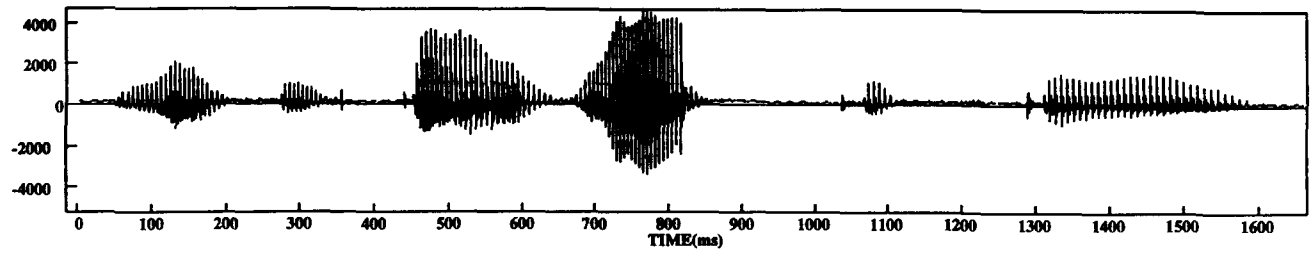

Figure 1. *Sound pressure signal for the sentence "Lisa's boy walked to school," spoken by a female talker. Syllable 4 was emphasized and has the greatest amplitude, and syllables 2 and 5 have reduced vowels and the least amplitude. The other syllables fall somewhere in between.*

formant bandwidth (B1) increases, the amplitude of the F1 peak is reduced, and SPL is reduced as $20 \log_{10} B1$. Finally, if the spectral tilt of the voicing source extends down to the first-formant region, SPL will be significantly reduced. The parameters B1 and source spectral tilt may be influenced by degree of glottal abduction, and thus glottal configuration can affect vowel amplitude.

3. EXPERIMENT

We collected data from four speakers (two females and two males), all adult native speakers of American English. The utterances recorded were reiterant versions of "Lisa's boy walked to school." In reiterant speech, a sequence of CV syllables is spoken with the same prosodic pattern as the intended sentence. This method allows one to study effects of prosody on speech production, while keeping parameters such as vowel quality constant. Thus, phonetic sources of variation are removed, making it easier to analyze data. For our experiment, syllables having full vowels were replaced by the syllable /pæ/ and syllables having reduced vowels were replaced by the syllable /pə/. The subjects were asked to speak as naturally as possible, and to place emphasis on either the first, third, fourth, or sixth syllable. Five tokens of each stress pattern were recorded.

Data were collected during two sessions. In the first session, the acoustic sound pressure was recorded in a sound-treated room. In the second session, oral pressure signals were recorded with a pneumotachographic mask, modified to include a pressure transducer (for details, see [5]). Sound pressure level was recorded simultaneously. Both the acoustic and aerodynamic signals were digitized and stored on a computer for further analysis.

Several parameters were extracted from the acoustic sound pressure signals. The amplitude of the first harmonic, H1, relative to those of the second and third formants (A2 and A3), was measured for each syllable. These measurements were made on spectra calculated using a 25.6 ms Hamming window for female speakers and a 30 ms window for males. Increases in H1-A2 and H1-A3 indicate increases in source spectral tilt, possibly due to increased A_g [6]. The other parameter extracted for each syllable was fundamental frequency, F0.

Figure 2 shows an example of the oral pressure signal and the corresponding speech and SPL signals that were collected. Subglottal pressure during vowel production was estimated from the oral pressure during the following /p/ closure [7]. SPL for each syllable was estimated as the maximum value during the syllable. All parameter values were averaged across tokens to obtain a mean value of each stress pattern for each syllable.

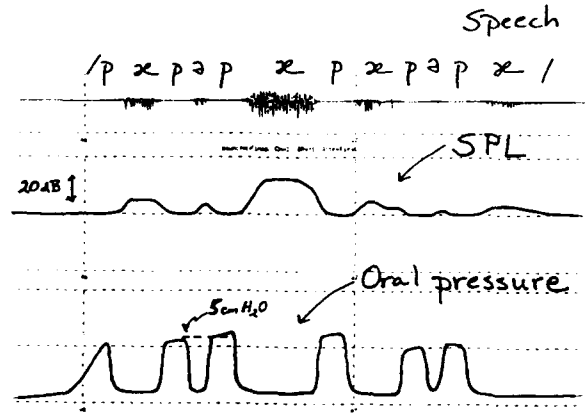

Figure 2. *An example of the data collected during the aerodynamic data recording session.*

4. RESULTS

We present results for the stress patterns in which either the third or fourth syllable has phrasal stress, and we further restrict analysis to syllables two through five of those sentences. The purpose of these constraints is to remove any end effects that might occur. These effects will be considered in future work.

Figure 3 shows graphs of SPL as a function of P_s for the two female speakers, F1 and F2. SPL has been corrected for changes in F0.[1] P_s is shown to the 3/2 power on a log scale. The solid points represent data from full vowels and the open points represent reduced vowels. The full vowel data are further labeled to indicate position in the stress pattern: because a syllable with phrasal stress carries a nuclear pitch accent, these syllables are labeled *nuclear*, while full vowels that occur before and after those with nuclear accent are labeled *prenuclear*[2] and *post-nuclear*, respectively.

[1] Note that SPL and F0 were measured from signals recorded in separate recording sessions, so the corrections are only estimates.

[2] This nomenclature does *not* mean that such syllables carry a prenuclear pitch accent; we only use it to indicate position of the syllable in a phrase.

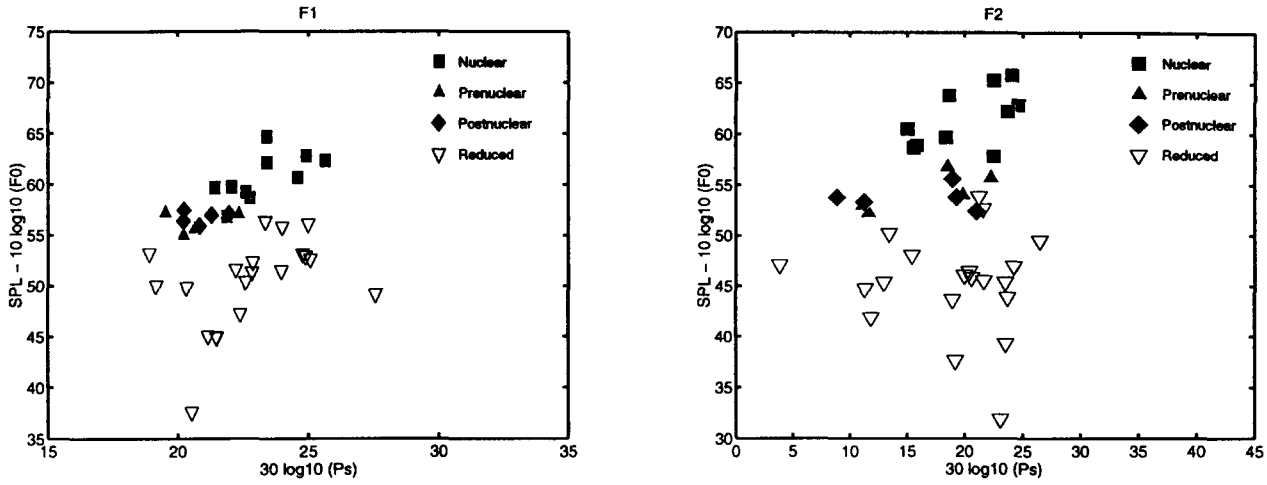

Figure 3. *Data for the female speakers, F1 and F2. Filled data points represent full vowels and open symbols represent reduced vowels.*

If only subglottal pressure was changing with SPL, we might expect the data to cluster along a line of slope 1 [3],[4], the reduced vowels being produced with low P_s and the full vowels with higher P_s. However, we see that both reduced and full vowels are produced with the same range of P_s, although full vowels have higher SPLs than reduced vowels produced with the same P_s; the average difference in SPL for a given value of P_s is about 8 dB for subject F1 and 11 dB for F2. Within the category of full vowel, subject F1 produces vowels in nuclear position with higher P_s and SPL than the other full vowels. A straight line fit to her full vowel data has a slope of 1.3, indicating that parameters other than P_s are changing as SPL increases. For example, the glottis may become more adducted, leading to reduced losses and thus increased SPL. The other female speaker, F2, also produces vowels in nuclear position with greater SPL than the other full vowels, but separation along the P_s dimension is not as great. A straight line fit to her full vowel data has a slope of 0.54.

Figure 4 shows the same type of graph for the male speakers, M1 and M2. Unlike the data for the female speakers, SPL does not increase with P_s. Thus, if earlier production data on sustained vowels or repeated syllables (e.g. [4]) is accurate, it is possible that the male speakers are actively making adjustments during production of more natural speech in order to prevent increases in SPL due to increasing P_s. The difference in average SPL for full and reduced vowels is only 3–4 dB for M1. His vowels in prenuclear and nuclear position are well separated along the SPL dimension from both reduced vowels and those in post-nuclear position along the SPL dimension. The difference in average SPL for full and reduced vowels is about 4 dB for subject M2, and his reduced vowels and those in both prenuclear and post-nuclear position are produced with less SPL than those in nuclear position.

In general, the subjects produced reduced vowels with the same range of P_s that they used to produce full vowels in nuclear-stress position, but with less SPL. The behavior for the other full vowels (prenuclear and post-nuclear) varies somewhat from speaker to speaker, but for the most part these vowels are also produced with the same range of P_s as are reduced vowels, but with greater SPL. How is it that speakers produce this vowel amplitude contrast without reducing P_s?

One possibility is that reduced vowels are produced with greater glottal abduction than full vowels, resulting in wider F1 bandwidths and greater source spectral tilt, and thus leading to reduced SPL for a given P_s. To explore this possibility we turned to the acoustic parameters H1-A2 and H1-A3, which can indicate changes in source spectral tilt, and thus are possible measures of the degree of glottal abduction. Values of the two acoustic parameters were averaged across stress patterns to obtain a mean value for each syllable type.

The results are shown in Fig. 5. Both H1-A2 and H1-A3 are significantly higher for reduced vowels than for full vowels, suggesting that glottal abduction is increased during the production of reduced vowels, relative to the abduction typical for full vowels. Averaged across the four speakers, H1-A2 is about 4 db for full vowels and about 13 dB for reduced vowels, a difference of 9 dB. H1-A3 is about 14 dB for full vowels and about 23 dB for reduced vowels, again a difference of about 9 dB. We also note that for three speakers the acoustic measures are higher for vowels in prenuclear and post-nuclear position than they are for those in nuclear position. Thus, glottal configuration may also play a role in the vowel amplitude contrast between these syllables. In general, then, our results indicate a tendency for glottal abduction to increase during production of vowels that are not in nuclear-stress position.

5. CONCLUSION

In summary, we have found that there are individual differences, and perhaps gender differences, in the relationship between SPL and P_s, and in the degree of vowel amplitude contrast between the four types of syllables examined (nuclear, prenuclear, post-nuclear, and reduced). Such differences may contribute to the perception of a speaker's individuality or gender. However, an overall trend can also be noted and may be useful for articulatory synthesis: P_s

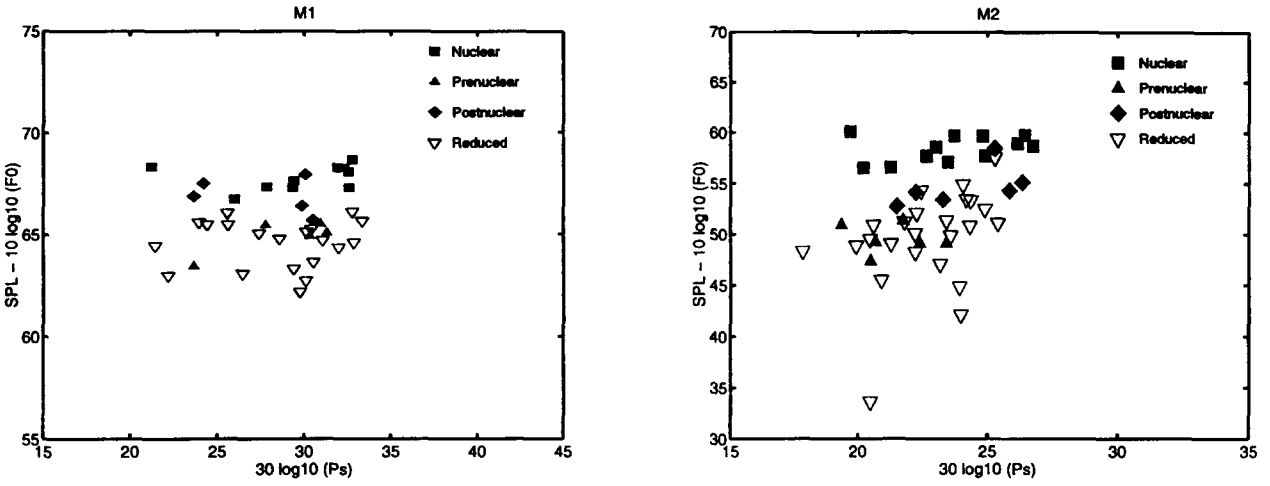

Figure 4. *Data for the male speakers, M1 and M2. Filled data points represent full vowels and open symbols represent reduced vowels.*

can be used to control vowel amplitude at sentence level and at main prominences (i.e. phrasal stress) [8], but vowel amplitude of reduced and non-nuclear full vowels may be manipulated by adjustment of the synthesizer parameters representing glottal configuration. More precise details of these adjustments remain to be studied. Finally, the variation of source spectral tilt between nuclear and non-nuclear full vowels has implications not only for speech synthesis, but also for speech recognition that is based on features such as cepstral coefficients, since some of these coefficients are sensitive to spectral tilt.

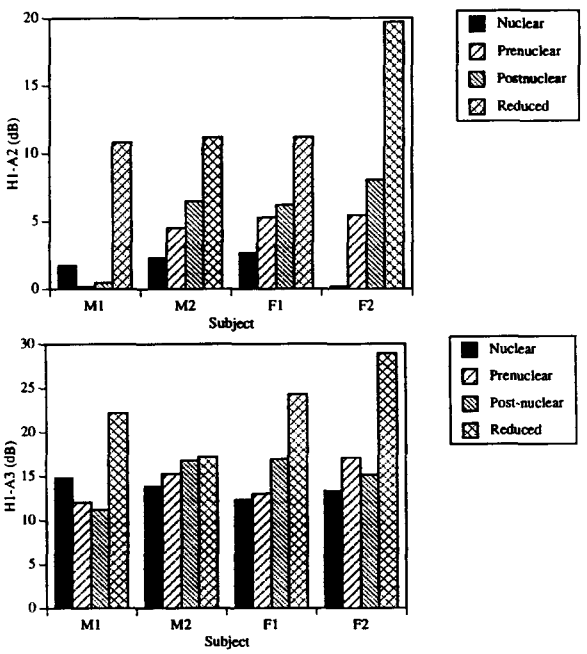

Figure 5. *The acoustic parameters H1-A2 and H1-A3 as functions of syllable position in the stress pattern of an utterance for each subject.*

REFERENCES

[1] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, pp. 820–857, 1990.

[2] K.N. Stevens and C.A. Bickley, "Constraints among parameters simplify control of Klatt formant synthesizer," *J. Phonetics*, vol. 19, pp. 161–174, 1991.

[3] G. Fant, "Preliminaries to analysis of the human voice source," *Speech Trans. Lab. Q. Prog. Stat. Rep.*, Stockholm, Royal Institute of Technology, vol. 4, pp. 1–27, 1982.

[4] S. Tanaka and W.J. Gould, "Relationships between vocal intensity and noninvasively obtained aerodynamic parameters in normal subjects," *J. Acoust. Soc. Am.*, vol. 73, pp. 1316–1321, 1983.

[5] E.B. Holmberg, R.E. Hillman, and J.S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice," *J. Acoust. Soc. Am.*, vol. 84, pp. 511–529, 1988.

[6] H.M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *J. Acoust. Soc. Am.*, in press.

[7] S. Hertegård, J. Gauffin, and P.-Å. Lindestad, (1994). "A comparison of subglottal and intraoral pressure measurements during phonation," *J. Voice*, vol. 9, pp. 149–155, 1994.

[8] J.E. Atkinson, "Correlation analysis of the physiological factors controlling fundamental voice frequency," *J. Acoust. Soc. Am.*, vol. 63, pp. 211–222, 1978.

Acknowledgements: Thanks to Jane Wozniak, Majid Zandipour, and Joe Perkell of MIT for their assistance with the aerodynamic recordings. I am also grateful to the subjects who participated in the experiment, to Alice Turk for her advice on prosody, and to Ken Stevens for many helpful discussions and for his encouragement and support.