TONE DISTRIBUTION AND ITS EFFECT ON SUBGLOTTAL PRESSURE DURING SPEECH

Helen M. Hanson[*]*, Janet Slifka*[*]*, Stefanie Shattuck-Hufnagel*[*]*, and James B. Kobler*[†]

[*]Speech Communication Group, MIT; [†]Massachusetts General Hospital, Boston
[*] {hanson,slifka,stef}@speech.mit.edu, [†] jkobler@partners.org

ABSTRACT

The current work is part of a project to characterize the subglottal pressure (P_s) contour associated with a spoken utterance in terms of the distribution of pitch accents and of phrase and boundary tones. It is found that the nuclear pitch accent does not define the start of the termination phase; the utterance offset is a better marker. Declination rate of the working phase and its relation to the phrase and boundary tones at utterance offset are found to vary among speakers. The results have implications for models of speech production, and for applications such as computer speech synthesis and recognition.

Keywords: Respiration, declination, termination, prosodic constituents

1. INTRODUCTION

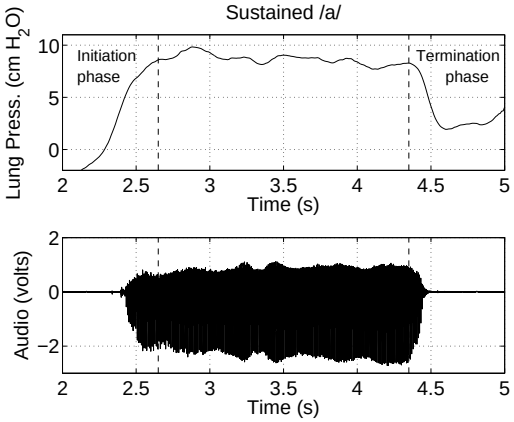
Some of the variability in the speech waveform is due to variation in subglottal pressure (P_s). There are three major sources of variation in P_s:

1. **Physiological constraints.** Speakers can hold only a finite amount of air in their lungs. Following exhalation of air during utterance production, there is a need to complete the exhalation and then rapidly inhale a new supply of air. If speech is produced during these times (before and after the working phase), when P_s is changing rapidly, the acoustic signal may be affected, particularly the amplitude [8].
2. **Segmental effects.** Variation of glottal and supraglottal impedances can result in intrinsic local dips or peaks in P_s [6].
3. **Prosodic effects.** Speakers may intentionally vary P_s when producing tones and boundaries (e.g., [1, 2, 5, 8, 9]), or increasing vocal effort.

Our longterm goal to model P_s variation for speech synthesis with an articulatory or quasi-articulatory speech synthesis system, such as HLsyn [3], will require addressing all three sources. In this paper, we focus on the third source of variation in P_s, the production of prosody.

Figure 1 illustrates the subglottal pressure contour for a sustained vowel. It is helpful to think of this exhalation portion as having three phases. The *initiation phase* is marked by a rapid increase in P_s, as the speaker begins to produce an utterance. It is followed by the *working phase*, during which P_s is relatively constant, or declining gradually (e.g., [1, 5], see Fig. 2 for an example), and the bulk of an utterance is produced. The working phase is followed by the *termination phase*, during which the utterance is completed and there occurs a rapid decrease in P_s.

Figure 1: Estimated lung pressure (upper panel) and acoustic waveform (second panel) for a sustained vowel /ɑ/ produced by a male speaker.



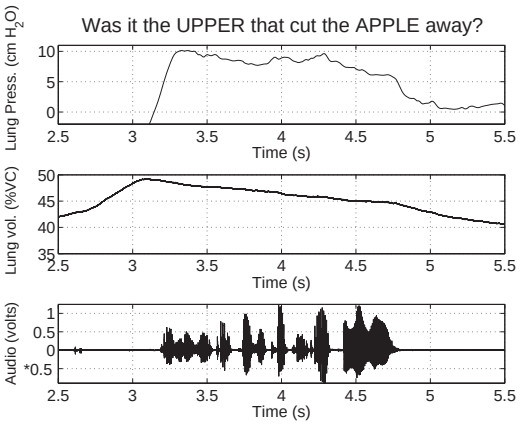
Our goal is to model aspects of the P_s contour in terms of the distribution of pitch accents, boundary tones, and phrase tones [7] of a speech utterance. More specifically, we would like to know (1) how to place the transitions between the three phases, and (2) how to set the slope of the working phase, depending on the type of pitch accents and phrase and boundary tones that occur in an utterance.

In this paper, we focus on the working and termination phases. We attempt to relate (1) the transition point between working and termination phases, and (2) the slope of the working phase, to the distribution of pitch accents, phrase tones, and boundary tones.

2. EXPERIMENTAL METHOD

We attempted to control the placement of pitch accents, phrase tones, and boundary tones in speech ut-

Figure 2: Estimated lung pressure (upper panel), lung volume (middle panel), and acoustic waveform (bottom panel) for an utterance produced by a male speaker.



terances by placing sentences in scripts designed to elicit the intonation contours listed in Table 1. The

Table 1: Intonation contours intended to be elicited from speakers. "TW" is an abbreviation for "target word," "PA" is an abbreviation for "pitch accent," and "–" (minus sign) indicates that the TW was not intended to carry a pitch accent. "CP" is an abbreviation for "carrier phrase."

TW1 PA	TW2 PA	Phrase and boundary tones	Carrier Phrase
H*	H*	L-L%	CP1, CP2
H*	–	L-L%	CP1
–	H*	L-L%	CP1
L*	L*	H-H%	CP3
L*	–	H-H%	CP3
–	L*	H-H%	CP3

target words were as follows: *pepper, tucker, capper, becker, ducker, gapper, apple, and upper*.[1] These were embedded in the following carrier phrases:

CP1 *It was the <TW1> that cut the <TW2> away.*

CP2 *The <TW1> cut the <TW2> away.*

CP3 *Was it the <TW1> that cut the <TW2> away?*

where TW is an abbreviation for "target word." An example script is as follows:

> Who cut the capper away?
> *Was it the apple that cut the capper away?*
> No, it wasn't the apple.
> *It was the pepper that cut the capper away.*

The lines in italics are the target utterances. In this example, the first sentence makes *capper* information that is "given" for the remainder of the script, in the hope that speakers will put the NPA on the first target words in the following utterances.

Data were collected at the Voice and Speech Laboratory at the Massachusetts Eye and Ear Infirmary (for details, see [8]). Six adult speakers of American English (three females and three males) served as subjects. The following signals were recorded:

1. Esophageal pressure (esophageal balloon)
2. Ribcage and abdomen cross-sectional areas (Respitrace)
3. Oral airflow (Rothenberg mask)
4. Oral pressure (*pae-pae-pae* strings only)
5. EGG
6. Audio

After recording the appropriate calibration data, the subjects were recorded producing (1) sustained vowels, (2) *pae-pae-pae* strings, and (3) the 50 scripts described above. Each script was recorded twice. An experienced ToBI labeler was present during the recordings to screen the utterances for target pitch contours. Subjects were asked to reproduce scripts produced with non-target pitch contours.

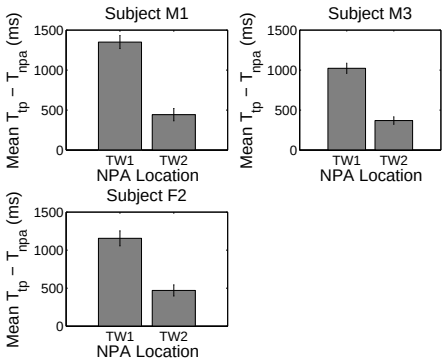
The signals were stored and processed using MATLAB. Following calibration of the signals, lung volume (LV) was derived from the two Respitrace signals. Esophageal pressure was then corrected for LV, to obtain the lung pressure (P_s) [4]. Utterances were checked to verify that at the beginning of an exhalation ($U_g = 0$), $-1.5 \leq P_s \leq 1.5$ cm H_2O [8]. Utterances that did not satisfy this criterion were discarded. Utterances were also discarded if subjects did not complete an exhale before producing them. All told, each subject produced about 125 utterances that can be used for analysis.

We are in the process of extracting F0 contours and labeling the audio and physiological signals for points such as utterance onset and offset, and peaks or valleys in F0. Table 2 summarizes the measures that will be discussed in this paper. Data for three subjects (one female, two males) have been analyzed and will be presented.

Table 2: Labels and measures made on the utterances.

Measure	Description
T_{npa}	Time of NPA peak or valley
T_{tp}	Beginning of termination phase
$T_{\text{utt-off}}$	Time of acoustic utterance offset
S	Slope of line fitted to working phase

Figure 3: The measure $T_{\text{tp}} - T_{\text{npa}}$ compared across position of the NPA.



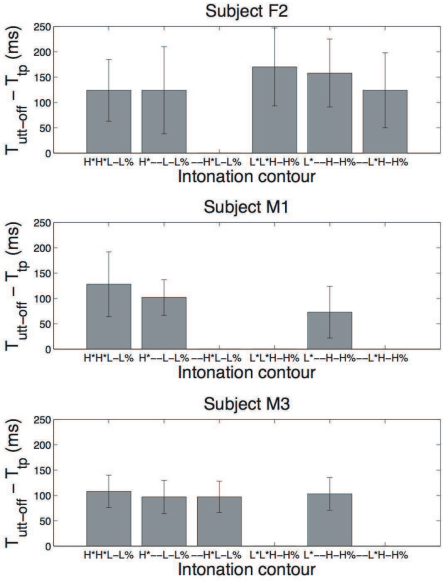
3. ANALYSIS AND RESULTS

The transition from the working phase to the termination phase can be difficult to identify [8]. Thus, our first task was to label the P_s contour of each utterance as either having a clear transition or not. An example of a clear transition can be seen in Fig. 1. The contour in Fig. 2 is an example of a contour that does not have a clear transition (e.g., is it at 4.25 s or at 4.75 s?). Because the measures analyzed in the remainder of this paper depend on having a well-defined T_{tp}, only the subset of utterances rated as having clear termination phases are used in the following analyses.

A first question was if the NPA or the utterance offset defines the start of the termination phase. We computed the time delay between the NPA (T_{npa}) and the beginning of the termination phase (T_{tp}), as well as that between the utterance offset ($T_{\text{utt-off}}$) and T_{tp}. The results for $T_{\text{tp}} - T_{\text{npa}}$ are displayed in Fig. 3, where the average time delay is compared for utterances with the NPA on TW1 and TW2. If the NPA marks the beginning of the termination phase, we would expect the time delay to be similar for the two possible locations of the NPA. However, it seems clear from this graph that the beginning of the termination phase is not influenced by the proximity of the nuclear pitch accent. While we have not completed statistical analyses, the lack of overlap of the error bars suggests that the measure $T_{\text{tp}} - T_{\text{npa}}$ is significantly different for the two locations of the NPA.

Figure 4 compares the results for $T_{\text{utt-off}} - T_{\text{tp}}$ across the six intonation contours. (A missing bar indicates that there were no more than two utterances with a given intonation contour that were judged to have a clear transition to the termination phase.) We see that utterance offset generally occurs after the termination phase begins. Although no statistical analysis has been done, the error bars

Figure 4: The measure $T_{\text{utt-off}} - T_{\text{tp}}$ compared across intonation contours.



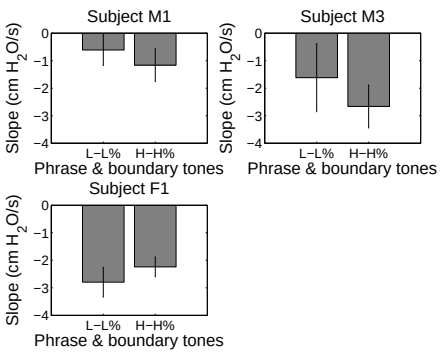
imply that the delay between T_{tp} and utterance offset is highly variable, and therefore it is difficult to use $T_{\text{utt-off}}$ to pinpoint the beginning of the termination phase. In addition, there does not seem to be an effect of tone distribution on this measure.

Next, we estimated the declination of P_s during the working phase. Straight lines were fit to the working phase of the P_s contour, and the slope S was taken as the estimated declination.

Our first question regarding the declination rate was whether S would be affected by the phrase and boundary tones. One might expect that P_s will decline more steeply for statements, because F0 decreases, while P_s will either rise or decline more gradually for questions because F0 increases (see e.g., [1, 5] for a discussion of the relation between F0 and P_s). Average slopes were computed for the two combinations of boundary and phrase tones used (L-L% and H-H%), and the results are displayed in Fig. 5. As can be seen, P_s tends to decline more rapidly during statements only for subject F1; the other two subjects show a tendency for a greater declination rate for the questions.

Furthermore, this figure illustrates that there are significant differences among the subjects in the degree of P_s declination. It is possible that normalization of our measures by sentence duration would result in more uniform results across speakers. Nonetheless, such differences in rate of declination could lead to "signature" variations in F0 and SPL that, combined with speaking rate, contribute to

Figure 5: Comparison of working-phase declination rates (slope) for phrases that end with L-L% and H-H% phrase-and-boundary-tone combinations.



speaker individuality. We will explore the effects of time normalization in future work.

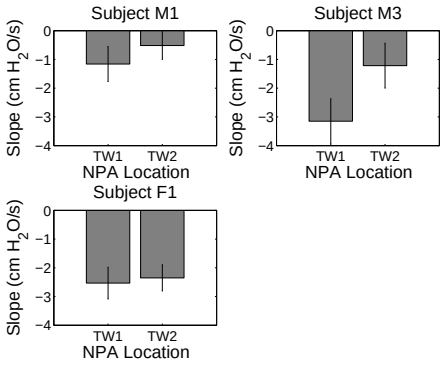
Next we computed the average slopes for the two possible positions of the NPA, with the expectation that P_s might decline more rapidly when the NPA is on TW1 than on TW2. The results are plotted in Fig. 6. Only data for subject M3 show the expected result. For subject F2, there appears to be no difference. Subject M1 shows some difference in the expected direction, but based on the degree of variation suggested by the error bars, this difference may not be statistically significant.

4. Conclusions

For the subset of data analyzed, we found that the NPA does not mark the beginning of the termination phase of the P_s contour. Utterance offset is a better marker, but it generally occurs after the beginning of the termination phase and the delay varies greatly across token and speaker. Declination of the working phase seems to be affected by the combination of phrase and boundary tones that occur at utterance offset, but this effect is again dependent on speaker. The declination is not as strongly affected by the distribution of pitch accents.

In future work, we will continue analysis of the entire database, in hope of resolving some of the less clear results. A critical task is to label the beginning of the termination phase for those utterances that do not have a clear transition between the working and termination phases. We are exploring various methods for doing so. A simple estimate would be to use $T_{\text{utt-off}}$ (see earlier discussion). Although that measure was found to be somewhat variable in relation to T_{tp}, it could prove to be adequate for some measures. By including more data, we can do detailed statistical analyses.

Figure 6: Comparison of working-phase declination rates (slope) for different positions of the NPA.



5. REFERENCES

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

[2] Hanson, H. M. 1997. Vowel amplitude variation during sentence production. In: *Proc. ICASSP-97* Munich, 1627–1630.

[3] Hanson, H. M., Stevens, K. N. 2002. A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *J. Acoust. Soc. Am.* 112, 1158–1182.

[4] Kunze, L. H. 1964. Evaluation of methods of estimating sub-glottal air pressure. *J. Speech Hear. Res.* 7, 151–164,

[5] Ladefoged, P. 1968. Linguistic aspects of respiratory phenomena. *Sound Production in Man*. New York: New York Academy of Sciences, 141–151.

[6] Ohala, J. J. 1990. Respiratory activity in speech. In: Hardcastle, W. J., Marchal, A. (eds.), *Speech Production and Speech Modelling*. Boston: Kluwer Academic Publishers, 23–53.

[7] Silverman, K. E. A., Beckman, M., Pitrelli, J. F., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. TOBI: A standard for labeling English prosody. In: *Proc. ICSLP* Banff, Canada 867–870.

[8] Slifka, J. 2000. *Respiratory Constraints at Prosodic Boundaries in Speech*. PhD thesis, MIT, Cambridge, MA. http://hdl.handle.net/1721.1/29184 visited 31-May-07.

[9] Slifka, J. 2003. Respiratory constraints on speech production: Starting an utterance. *J. Acoust. Soc. Am.* 114, 3343–3353.

[1] The stop consonants may introduce segmental effects on P_s, but we assume that these small, local effects will not affect the more global aspects addressed here.