Stop Consonant Production: An Articulation and Acoustic Study

by

Kelly L. Poort

B.S., Electrical Engineering, University of Iowa
B.S., Biomedical Engineering, University of Iowa
(1991)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1995

Author ....
Department of Electrical Engineering and Computer Science
August 11, 1995

Certified by ....
Kenneth N. Stevens
Clarence J. LeBel Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Stop Consonant Production:

An Articulation and Acoustic Study

by

Kelly L. Poort

Submitted to the Department of Electrical Engineering and Computer Science
on August 11, 1995, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

This thesis investigated the production of the labial and alveolar voiceless stop consonants through examination of experimentally-recorded articulation and acoustic data. Articulator movements, as measured by transducers placed on the individual articulators, were recorded by an electro-magnetic midsagittal articulometer (EMMA) system. The acoustic signal was recorded simultaneously. Three male speakers, of normal speech and hearing, spoke single-syllable /CVt/ sequences occurring naturally in the English language. The tokens were composed of the voiceless labial or alveolar stop consonant C followed by one of the vowels /ɑ, i/ and the consonant /t/, and were placed in a carrier phrase. Half the tokens were preceded by the fricative consonant /s/. For comparison purposes, the remaining stop and nasal consonants were also recorded.

A processing technique was developed to average the articulator displacement waveforms for a given utterance and speaker. The technique accounted for variations in speaking rate by linearly time warping the waveforms based on the timing of events observed in the acoustic signal, i.e. the stop release. Analysis of average displacement waveforms and average maximum downward velocities following the stop release yielded several observations: (1) The jaw, constrained to remain in a high position until the end of /s/, compensates by moving rapidly downward at a rate 1.4 cm^2/sec higher, on average, than utterances not containing /s/; (2) Further evidence of a correlation between larger articulator displacements and faster rates of movement is noted upon comparing utterances containing /ɑ/ vs. /i/; (3) Examining the lower lip movement for /p/ in pot, Subject 2 has a strategy of making a larger movement with a faster maximum velocity following release, than the other two speakers, yet takes more time to produce the stop, possibly indicating incomplete compensation for the larger movement; (4) A bandwidth of approximately 3 - 5 Hz exists for "cyclic" articulatory movements produced in speech, such the lower jaw moving from the steady-state portion of /e/ in say to the steady-state portion of /ɑ/ in pot, in "Say pot again."; and (5) The primary articulators also exhibit a single period of "cyclic" movement when producing a stop, with period durations of 200 - 300 msec.

Linear estimates of the rates of labial and alveolar constriction cross-sectional area increase following the release were determined from the articulation data, with the aid of models and the acoustic data. First, the frication noise burst was measured from the acoustic data. The burst duration was shorter for utterances containing an unaspirated /p/ (7 - 10 msec) than an unaspirated /t/ (11 - 16 msec). Second, an initial approximation to the constriction cross-sectional area change with time following the release was derived from the articulation data. An ellipse was used to model the labial constriction and the segment of a circle to model the alveolar constriction. A linear rate of area increase was estimated from the first 10 msec of the modeled cross-sectional area. Third, the linear rate was used in a low-frequency circuit model of the average pressures and airflows in the vocal tract during stop production. Fourth, the airflow through the constriction, an output of the circuit model, was used in a model of the frication noise burst generated at the constriction. Fifth, the duration of the modeled noise burst was compared to the measured noise burst duration. Lastly, the modeled cross-sectional area and associated linear rate were adjusted, and the above process beginning with the circuit model repeated, until the linear rate produced a modeled noise burst equal in duration to the measured burst. The resulting linear rates of area increase were in the range 35 - 50 cm^2/sec for the voiceless labial stop, and 20 - 30 cm^2/sec for the voiceless alveolar stop.

The linear rate of constriction cross-sectional area increase was incorporated into a high-frequency model of the vocal tract. Given the cross-sectional area along the vocal tract length, the wave equation was solved for the $F1$ transition following the stop release. $F1$ values were also measured from the acoustic data. The estimated $F1$ transition supplemented the acoustic data during the frication noise burst time period when $F1$ was not excited by a vocal-tract source. The computed $F1$ transition agreed well with the measured $F1$ values in the vowel, indicating that the approach used to estimate the linear rate of constriction cross-sectional area increase following the release was reasonable. It was also shown that an exponential curve fit to the acoustic data does not consistently provide a good estimate of the $F1$ transition.

Thesis Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering

Dedication

In memory of,

two whose time with me was altogether too short,
yet who made such a difference in my life,

Elizabeth (Betsy) A. Dunlap, my cousin,
who helped me realize family and friends are the most important things in life,

and

Dr. Paul D. Scholz, Associate Dean, College of Engineering, University of Iowa,
who, through his belief and pride in me, encouraged me to reach for my goals.

Acknowledgments

My deepest appreciation goes to my thesis advisor, Professor Ken Stevens, for his guidance, never-ending patience, and particularly for sharing with me his enthusiasm and love for speech.

I would also like to thank Joe Perkell, Melanie Matthies and Mario Svirsky for the opportunity to use the EMMA system and for their willingness to answer my numerous questions.

A special thank you to Corine Bickley, who has been a constant source of much-needed encouragement, to Lorin Wilde for inspiring me by her example, and to Jane Wozniak, for providing a friendly ear on countless occasions. Marilyn Chen and Jeff Kuo have proven to be invaluable resources for the HST program, providing me not only with advice, but class notes as well.

The talks I have given throughout the years have benefited greatly from the good advice of Stefanie Shattuck-Hufnagel. I especially thank her for making sure that all my boundary tones didn't rise at the end of my intonational phrases!

I am especially grateful to Professor Martha Gray, my academic advisor, for her unfailing belief in me and for calming my qualifier worries during our once-a-semester conversations.

My parents, grandparents, sister and brother-in-law (and my nephew, the newest family member!) have all been a steadfast source of support and encouragement.

Finally, my most heartfelt appreciation goes to my husband, Brian Sperry, without whom this thesis would not have been possible. You have been my everything, Brian, from my strongest supporter to my favorite typist. I am so delighted we are sharing the unique and wonderful "MIT Experience"!

This research was supported in part by grant DC00075 from the National Institutes of Health.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Statement of Purpose

The speech sounds of language can be divided into two categories, vowels and consonants. Vowels are produced with no constriction in the vocal tract, whereas consonants are formed by making a constriction at some point between the glottis and the lips. Consonants range from those creating no pressure buildup in the vocal tract (sonorants) to those which narrowly constrict or completely close the vocal tract, resulting in substantial pressure buildup (obstruents). The majority of speech sounds, including vowels, sonorants and fricatives (one type of obstruent), contain a steady-state time period in which sound is continually generated. The only two classes of speech sounds that contain no steady state are the remaining types of obstruents, i.e., stops and affricates. These two obstruent types are characterized by acoustic properties that change rapidly with the progression of time; thus, they must be described by sampling the acoustic properties at a series of time intervals. Historically, it has been easier to study the production of sounds containing a steady state, due to the low frame rate of many of the available measurement techniques. As a result, there is not much information available about the detailed production and acoustics of stop consonants, the focus of the present study.

The formation of a stop consonant normally consists of three consecutive phases: the onset of closure, when one articulator is approaching the other; the closure, when the articulators are held together, completely obstructing the airflow and creating a pressure buildup behind the constriction; and the offset of closure, when the articulators are moving apart again (Henton, Ladefoged and Maddieson, 1992). In the English language, stop consonants are formed by making the constriction at a point between the lips and the velum. The stops can be divided into two types, voiced (/b, d, g/) and voiceless (/p, t, k/), depending upon whether the vocal folds are vibrating throughout most of the second and third phases or not, respectively. When there is voicing, the glottal source excites the vocal-tract formants, resulting in radiated sound which can reveal many aspects of the articulator movements throughout all three phases. For voiceless stops, however, there is no periodic excitation of the formants during closure and only noise excitation of the rapidly-changing vocal tract at the beginning of the offset phase. Consequently, it is more difficult to determine articulator movements solely from the acoustic waveform. The present study proposes combining articulatory information for the voiceless stops, derived from recordings of the physical movements of the articulators, with simultaneous acoustic recordings. The coupling of articulatory and acoustic information should provide further understanding of stop-consonant production, leading to knowledge of the sequence and timing of articulator movements as well as the resultant acoustics. In addition, it is expected that examination of the articulation information will lead to improved interpretation of the acoustic signal, such that in the future only the acoustic waveform is required to determine numerous important aspects of the vocal-tract configuration for stops.

Some applications for the results of the study include speech recognition, speech synthesis, and diagnosis and remediation of disordered speech production. For recognition and synthesis, improved knowledge of the articulator movements and their acoustic correlates would assist in establishing parameter ranges for models of stop production and recognition. Refining and expanding the models will also contribute to theories of control and coordination of the articulatory structures during speech production. The parameter ranges could then serve as a baseline against which disordered speech production is evaluated. The ability to determine most, if not all, of the articulators' movements from the acoustic signal would be a useful clinical aid in determining the ways in which articulators are moving incorrectly for a person with disordered speech.

1.2 Literature Survey

A considerable volume of research has been dedicated to determining the acoustic cues for correct perceptual identification of stop consonants (Liberman, Delattre and Cooper, 1958; Lisker and Abramson, 1964; Lisker and Abramson, 1970; Stevens and Klatt, 1974; Dorman and Raphael, 1980; Ohde and Stevens, 1983; Ohde, 1984; Nossair and Zahorian, 1991, and many others). By contrast, notably fewer studies have utilized acoustics to indirectly infer the production of stop consonants. The primary approach for determining aspects of stop production has been to attempt to directly measure the movement of the articulators with various types of measurement devices. Consequently, stop production has been examined almost entirely through the use of a limited number of articulation studies. The disparity between the quantity of perception and production studies is primarily attributed to the frame rate constraint mentioned earlier, as well as the limited availability of articulation measurement devices. The following two sections give a chronological summary of previous stop-consonant production research in the articulation and acoustic areas.

1.2.1 Articulation Experiments

Some of the first quantitative studies of the articulatory motions involved in stop-consonant production were performed using x-ray cineradiography in the 1960's - 1970's. Perkell (1969) used an x-ray cineradiographic system to obtain data on the relative articulator positions for the voiceless stops in /hə'Cɛ/ sequences, spoken by one speaker and sampled at a frame rate of about 40 to 60 Hz. Data were obtained for the movements of the primary articulator and jaw during the transitions into and out of closure. The primary articulator is anatomically anchored to the lower mandible and forms the complete closure in the vocal tract. For /p/ the primary articulator is the lower lip, for /t/ the tongue tip and for /k/ the tongue body. A secondary articulator is any other articulator which assists in the formation of the closure but does not make the actual constriction. A stop has only one primary articulator, but may have several secondary articulators. For each of the three voiceless stop consonants, the jaw functions as a secondary articulator. In this particular study, the movement of the lower lip during /p/ production was determined by measuring the vertical distance between the lowermost part of the upper lip contour and the uppermost part of the lower lip contour. This distance is referred to as the "height of the lip aperture". The movements of the primary articulators for /t/ and /k/ were measured along axes fixed to the maxilla. The velocities of the articulators at the time of constriction opening, derived from the data of Perkell (1969), are summarized in Table 1.1.

Stop Consonant	Jaw Opening Velocity (cm/sec)	Primary Articulator Opening Velocity (cm/sec)
/p/	6	20
/t/	4	8
/k/	5	12

Table 1.1: Jaw and primary articulator velocities at time of constriction opening for /p, t, k/ (Derived from Perkell, 1969).

The tabulated values are the average velocities for the time period surrounding the transition into the closure offset phase. The time period over which the velocity is calculated varies from approximately 75 to 150 msec. The juncture of the closure and offset phases, when the constriction is initially starting to open, is referred to as the stop-consonant release. Average rates of articulator movement near the time of closure onset could be similarly derived.

A cineradiographic study by Kuehn and Moll (1976) consisted of /VCVC/ utterances, where C was a voiceless stop and V one of the vowels /i, ɑ, u/. The vowel was varied to determine the effect of phonetic context on stop-consonant production. The utterances were spoken by five subjects and recorded at a frame rate of 150 Hz. The movements of the primary articulators were measured within two sets of axes, one set fixed to maxillary structures and the other fixed to mandibular structures. Some of the across-speaker conclusions of the study include: (1) The velocities of the primary articulators in the utterances containing the low vowel /ɑ/ are generally greater than those of the high vowels /i, u/, regardless of the consonants involved. These velocity differences due to vowel context were found to be closely related to corresponding differences in extent of articulator displacement. In general, the larger the displacement, the faster the velocity. Although previous studies had also made this observation (including Perkell, 1969), Kuehn and Moll determined that the velocity-displacement relationship holds across speakers and articulators; (2) Faster velocities are associated with /p, t/ than with /k/; and (3) The primary articulator velocity as the constriction closes is notably faster than at the consonant release. The latter two across-speaker conclusions are evident in Table 1.2 (averaged across all vowels and speakers).

Stop Consonant	Primary Articulator Closing Velocity (cm/sec)	Primary Articulator Opening Velocity (cm/sec)
/p/	22	16
/t/	26	19
/k/	14	12

Table 1.2: Primary articulator velocities at time of constriction closure and opening for /p, t, k/ (Kuehn and Moll, 1976).

These latter two conclusions hold even when the data are adjusted to account for the dependence of velocity on the displacement of the nearby vowels. A final observation of the study found that velocity differences between speakers may be primarily accounted for by vocal tract size. Subjects with larger vocal tracts must move their articulators greater distances between sounds in order to correctly produce them. Since the Kuehn and Moll study determined that the time it takes to produce a given sound remains relatively constant across speakers, subjects with larger vocal tracts must be moving their articulators faster. (It is possible that sex may be a confounding factor in this observation, since female speakers had smaller jaw sizes than males).

Gay (1977) utilized a cineradiographic system to examine articulatory movements in /CVCVC/ sequences. The initial and final consonants were fixed (/k/ and /p/, respectively), and the medial /VCV/ contained the vowels /i, ɑ, u/ and the three voiceless stops in all possible combinations. The utterances were placed in a carrier phrase, spoken by two subjects and sampled at 60 Hz. An important outcome of the study was an observation of the presence of coarticulation in stop-consonant production. Results revealed that articulatory movements in anticipation of the second vowel began during the closure period of the intervocalic stop consonant. A second study, performed by Borden and Gay (1979), further investigated the coarticulatory effects. Three speakers recorded utterances composed of /sCɑpə/, with C one of the voiceless stops, at a frame rate of 60 Hz. Borden and Gay found that the high lower lip position necessary for /p/ closure and the high tongue tip position necessary for /t/ closure each placed constraints upon the jaw, forcing it to remain high during the entire production of /s/ preceding /p/ or /t/. For /k/, however, the jaw began to lower during the /s/, indicating that the jaw does not need to be elevated to the samc degree to maintain closure of the airway in /k/ as in /p, t/. The study also determined that, in general, only the portion of the tongue primarily involved in the production of a stop, like the tongue tip for /t/ and the tongue body for /k/, has restrictions placed on its movements. The remainder of the tongue is somewhat free to anticipate the position of the following vowel. This finding suggests that the tongue is composed of parts able to move relatively independently of one another. Lastly, /s/ was noted to have a shorter duration prior to /p/ than prior to /t/ or /k/, probably because the lips are not involved in any conflicting gesture and are free to close. In contrast, the tongue is involved with the /s/ constriction, causing movement toward closure to be delayed and gradual.

In the x-ray cineradiography studies, the frame rate varied from 40 to 150 Hz, which is low enough that the actual opening and closing articulator velocities may be misrepresented by measured velocities which are too slow (particularly in the studies with 40 - 60 Hz frame rates). The low rate makes it especially difficult to determine the movements of the articulators during the 5 - 10 msec interval immediately following the release, when the articulators are moving very rapidly. Within that brief interval there may be at most only two sample points for the 150 Hz rate, and, on average, only one (or none, for the 40 Hz rate) will fall in the region. In addition, the long exposure time per frame may result in blurring and smoothing of the data in the spatial domain. Finally, only indirect information can be obtained on the constriction cross-sectional area due to the lateral view of the speaker during filming; however, the length of the constriction is readily determined.

By the beginning of the 1970's a strain gage system able to record superior-inferior movements of the lips and jaw had been developed. The system was found to be capable of faithfully following and recording the highest rates of articulator movement physically possible. Sussman, MacNeilage, and Hanson (1973) utilized the system to observe labial and mandibular movements during the production of bilabial consonants. Five subjects spoke /VCV/ utterances composed of all conceivable combinations of the vowels /i, ɛ, æ/ surrounding the bilabials /p, b, m/ and the voiceless stop /t/, included as a control. All utterances were spoken in a carrier phrase. Several results regarding timing, displacement and velocity ensued from the study. First, it was discovered that the upper lip started to lower for the bilabials approximately midway through the first vowel and reached its maximum lowering point coincident with the end of the first vowel. The lower lip started elevation toward closure in the middle of the first vowel, reached its highest elevation at the midpoint of the C closure phase, and completed its lowering for the second vowel in the vicinity of the onset of the second vowel. Second, the jaw was observed to lower most during the vowels surrounding /m/ (11.67 mm) then /b/ (11.57 mm) and lastly /p/ (11.25 mm), averaged across all vowels. A similar trend was seen for the lower lip. Larger jaw and lower lip displacements were seen for the lower vowels. Third, velocities of the various articulators were reported. The jaw and lower lip velocities for the bilabial consonants are summarized in Table 1.3. A few trends can be noted from the data: (1) the jaw shows a slightly faster closing than opening velocity on average, yet the lower lip shows a strong opposite trend; (2) the jaw lowering velocity is slowest after /p/, regardless of the following vowel; (3) the jaw, for both opening and closing, reveals a consistent increase in velocity as the vowel changes from high to low, evidence of a trend toward increasing velocity with increasing distance to be traveled; (4) the

Bilabial Cons. & Artic.	Closing Velocity (cm/sec)				Opening Velocity (cm/sec)			
	i → C	ɛ → C	æ → C	Mean	C → i	C → ɛ	C → æ	Mean
/p/-J	1.5	5.9	8.3	5.2	2.6	4.4	6.3	4.4
/p/-LL	12.6	12.2	12.0	12.2	16.3	17.6	18.9	17.6
/b/-J	2.2	5.3	7.5	5.0	2.8	5.0	6.8	4.8
/b/-LL	13.7	12.4	12.6	12.9	16.0	19.2	18.6	18.0
/m/-J	1.9	4.6	6.6	4.4	2.8	4.7	6.6	4.7
/m/-LL	13.0	12.6	12.8	12.8	18.2	19.1	18.4	18.6

Table 1.3: Jaw and lower lip closing and opening velocities for the bilabial consonants /p, b, m/ preceding and following the vowels /i, ɛ, æ/. Velocity values are averaged across six subjects and rounded to the nearest tenth cm/sec. Articulators: J = Jaw, LL = Lower Lip. (Based on Sussman, MacNeilage, and Hanson (1973), Table 1.)

lower lip velocities are, for the most part, not significantly different from one vowel to another; and (5) the velocities of the upper lip (not shown in the table) and of the sum of the jaw and lower lip follow the velocity hierarchy /p/ > /b/ > /m/ at the closure and /p/ = /b/ = /m/ at the release. Some attempt was made to relate these discoveries to the aerodynamic events of bilabial production. In general, the consonant with the least amount of pressure buildup, /m/, had the least and slowest amount of articulatory adjustment in preparation for production, and the stop with the highest intraoral air pressure, /p/, had the fastest and most pronounced articulatory movements. As a final note, a coarticulation effect of the first vowel on the jaw height of the second vowel was found to exist, influencing /m/ the most, since the opening gesture was the largest and fastest, and /t/ the least, since the jaw, being the most involved in forming the closure, was more constrained.

During the late 1970's and early 1980's, the strain gage system was improved in several ways (Müller and Abbs, 1979; Barlow, Cole and Abbs, 1983). Gracco (1994) utilized the refined system to study the production of bilabial stop and nasal consonants. Four subjects spoke utterances consisting of /sVCæp/ where V was one of /i, æ/ and C was a bilabial /p, b, m/. The sampling rate was 500 Hz (400 Hz for one subject). To eliminate high frequency noise, the movement signals were filtered by a low-pass, two-pole, zero phase lag, 20 Hz cutoff Butterworth filter. (Of note, Müller, Abbs, Kennedy and Larson, at an American Speech and Hearing Association meeting in 1977, presented spectral analyses of articulatory movements which indicated that very little frequency information lies above 10 Hz (Linville, 1982)). A unit of measure referred to as the "movement cycle" was employed in the study. A movement cycle was defined to start at the initiation of the opening movement for V and end at the termination of the closing movement for C, as measured from the lip aperture signal, a signal derived from a combination of the movements of the upper lip, lower lip and jaw. Movement cycle initiation and termination times were determined by the locations of the velocity zero crossings. The findings of the study include the following: (1) averaged across all speakers, the cycle duration was shorter for /iC/ than /æC/ context (191.8 vs. 227.3 msec) and shorter when C was /p/ than either /b/ or /m/ (203.3 vs. 210.2 vs. 215.1 msec, respectively); (2) the vertical lip opening for /s/, measured from the lip aperture signal, was invariant to subsequent phonetic context; (3) the jaw was the articulator solely responsible for the differences observed in lip openings of the vowels in the /VC/ context; (4) the lower lip closing velocity for C was found to be significantly higher for /p/ than /b/ or /m/ in the /æC/ context for three of the subjects; and (5) during the bilabial closure offset phase, the lower lip opening velocity is always lowest for /p/ and highest for /m/. Gracco also hypothesized that a potential explanation for voiced/voiceless differences originates in the relative timing of opening and closing gestures. For /p/, the closing action was initiated earlier and the opening gesture later than either /b/ or /m/, consistent with acoustic studies showing a longer acoustic closure duration for voiceless sounds. This trend may be a result of the need to accommodate the laryngeal adjustment associated with devoicing of the /p/.

Additional articulation studies performed over the two decades spanning 1970 to 1990 determined the maximum rates of articulator movements for both speech and nonspeech gestures. Nelson, Perkell and Westbury (1984) performed a strain gage study which found peak velocities of a single, nonspeech mandible movement (as measured at the lower incisors) to range upwards of 60 cm/sec. A maximum frequency of 7 Hz (corresponding to a period of approximately 150 msec) for alternating movements of the jaw in repeated-syllable speech was also discovered. Peak velocities of mandible movement are in the range 10 - 20 cm/sec (Nelson et al., 1984; Kent and Moll, 1972; Beckman and Edwards, 1993). Smith and Gartenberg (1984) utilized a strain gage and obtained average peak mandibular velocities in the range of 6 - 8.5 cm/sec during speech production. Smith and McLean-Muse (1986) also used a strain gage and discovered average peak velocities for the upper lip of 3 cm/sec, lower lip 5.5 cm/sec, and jaw 6 cm/sec in speech. Peak velocities of movement of the body of the lower lip immediately following release of a labial stop consonant are estimated to be about 25 cm/sec or less (Fujimura, 1961b; Sussman et al., 1973). Kent and Moll (1972) used cinefluorography, frame rate 150 Hz, to detect tongue tip peak velocities of 20 - 40 cm/sec during the production of /d/ in a VCV environment, with the faster rates for the implosion phase of the stop. These estimates are consistent with Kuehn and Moll (1976) for the consonant /t/. Ostry and Munhall (1985) employed a computerized pulsed ultrasound system to find tongue peak velocities closer to 8 - 13 cm/sec in speech. For a velar stop, the maximum velocity of the tongue surfaces in a direction normal to the palatal surface is about 20 cm/sec (Kent and Moll, 1972). Kuehn and Moll (1976) report a tongue body peak velocity of 10 - 15 cm/sec, averaged over several speakers.

Drawbacks to a strain gage system include recording only the lip and jaw articulator movements, the need for careful placement of the jaw transducer to avoid recording facial muscle movements, and the inability to directly determine either the length or the cross-sectional area of the labial constriction. In addition, there is the possibility that the response time of the transducers is too slow to record the most rapid movements of the articulators, such as those occurring at the consonant closure and release. However, average maximum velocities of the lips and jaw recorded by the strain gage system during speech are in the range of 3 - 8 cm/sec, well below the maximum velocity that the strain gage system is capable of recording, which is at least 60 cm/sec. The duration of the response time is apparently short enough to record even the most rapid articulator movements made during speech production.

In the 1980's to 1990's a new device, the x-ray microbeam system, was utilized to study stop-consonant production. Edwards (1985) employed the system to study the production of the voiceless stop /t/ in various phonetic environments. The utterances were of the form /V_1tV_2/, with V_1 one of /i, ɑ/ and V_2 one of /i, ɑ, æ/, spoken within a carrier phrase by one speaker. The frame rate was approximately 145 Hz. Edwards observed that tongue movement toward V_2 was initiated during /t/ consonant closure. Evidence also supported an interaction between the tongue and jaw that attempted to maintain consistent production of /t/ by compensating for some of the anticipatory coarticulation influences. A second x-ray microbeam stop-consonant study was performed by Macchi (1988). The utterances were /ipi/ or /ɑpɑ/, placed in carrier phrases and spoken by two speakers. The frame rate varied from 145 to 183 Hz. The jaw was noted to have a low position for the low vowel and a high position for the high vowel; however, the lip pellet position for /p/ was relatively invariant between the two vowel environments. A third similar study, by Papcun et al. (1992), recorded the vertical movements of the lower lip, tongue tip, and tongue dorsum of three speakers saying the voiced and voiceless stops in repeated /Cə/ syllables. Sampling rates for the pellets were 90 Hz for the lower lip and tongue dorsum, 180 Hz for the tongue tip. Papcun et al., observed that the movements of the primary articulators span a larger range and are less variable than the movements of the noncritical articulators in each utterance. A noncritical articulator is any articulator which is not a primary articulator, including, but not limited to, all secondary articulators. Because the primary articulator is responsible for the articulation of each consonant, its movements are more prescribed and constrained. Noncritical articulators, less constrained by segmental considerations, are freer to vary, and may have less acoustic effect or at least an acoustic effect that is perceptually less important. One possible explanation for inter- and/or intra-speaker variability is the variation in movements of the noncritical articulators. It should be noted that the x-ray microbeam system has the disadvantage of tracking only a set of point movements in the midsagittal plane, and consequently cannot supply detailed information about the vocal-tract shape, for example at a constriction. Also, the frame rates used in the three studies listed are somewhat slow, making it difficult to study rapid movements occurring over short periods of time. An advantage of the x-ray microbeam system is the ability to obtain fairly accurate measurements of tongue movements, a difficult or impossible task with most available devices.

In an attempt to explain and integrate many of the observations of articulatory movements, a task dynamic model was proposed by Saltzman (1986). The model considers movement of the individual articulators to be subordinate to the combined movement of the contributing parts used to produce a particular sound. In other words, the articulators do not move independently of one another, but rather their movements are related according to the task to be performed. The model suggests the presence of a higher-level speech motor plan, guiding the individual articulators to work as a unit in the production of each sound. The nervous system is believed to control the coordinative requirements of all the articulators (Gracco, 1990; 1991). Perkell (1994) points out that, in addition to the direct control provided by the nervous system, there is also feedback control via the auditory and orosensory systems. The task dynamic model can be applied to the production of stop consonants, attempting to account for the coordination, sequence, timing and rates of the articulators, as well as the effects of coarticulation. Gracco and Löfqvist (1993) utilized utterances containing /V_1pV_2/ to examine the articulator coordination upon entering and exiting the consonantal closure phase. They found that the lip, jaw and laryngeal movements during the act of closing had a greater degree of coupling than movements of the same articulators during the act of opening. The lower correlation among the articulators at the opening is attributed to the lips still being involved in the consonant sound while the jaw and larynx become functionally decoupled from the consonant and are directly involved in the production of the following vowel. Gracco and Löfqvist viewed the two actions, closing and opening, to reflect two important but distinct characteristics of speech production. For the closing, examination of the relative timing focused on the production of a single speech unit (the phoneme /p/ in this case) and the consistent timing of the movements represented the coordination of multiple speech articulators within a specific speech sound. For the opening, examination of the relative timing focused on a transition region between two contiguous phonemes, a consonant and a vowel. Consequently, the study examined both multiarticulator coordination within a speech production unit (the closing, within a single phoneme) and the sequencing of such units into larger aggregates (the opening, forming part of a syllable).

1.2.2 Acoustic Experiments

Acoustic analysis with the intent to investigate production, rather than perception, of stop consonants is exceedingly rare in the literature. No studies were located which utilized solely the acoustic signal to analyze stop-consonant production and only two studies were identified which analyzed both acoustic and articulatory experimental data. These two studies examined the relationship between the articulatory movement data and the vocal-tract resonances of the output speech signal. A third, recent study employs perceptual judgment of synthesized speech to study one specific aspect of stop production. The idea of utilizing acoustics to discover more about stop-consonant production is apparently relatively unique, although not new, given that the first two studies were performed in the early 1960's.

The first acoustic study was conducted by Fujimura (1961a) and has never been published. The purpose of the study was to determine the effects of vowel context on the articulation of stop consonants. Two-syllable English words containing /V_1CV_2/ were examined, where C was /t/ or /k/ and the vowels V_1 and V_2 were chosen to have labial articulations significantly different from one another. Example utterances include <u>veto</u>, <u>echo</u> and <u>okay</u>. A stroboscopic motion picture system was employed. The utterances were spoken by three subjects, and the stroboscopic illumination rate was 60 Hz. The study attempted to relate the articulator movements of the lips and jaw to the transitions of the first and second formant frequencies, $F1$ and $F2$. The cross-sectional area of the lip opening was estimated from measurements of the lip opening height and width. A method was developed, based on Stevens and House (1956), to estimate $F1$ and $F2$ from the lip opening cross-sectional area. During the closure interval, in the absence of excitation, $F1$ was set to zero, based on the assumption that the vocal-tract walls are rigid. (It should be noted that studies performed in the 1970's determined that the vocal-tract walls have a nonzero, finite impedance. Consequently, $F1$ during the closure interval is more accurately set to a minimum value of 180 Hz. Detailed discussion of the theoretical and experimental determination of the value of $F1$ during the closure interval appears in Chapter 2, Section 2.2.) A change in $F2$ from about 800 to 1600 Hz was calculated to occur during the closure period of /k/ in the word okay, corresponding to a measured change in the lip aperture from 0.3 cm^2 at the initial portion of the stop closure to 1.4 cm^2 by the end of the closure interval. During the noise production between the release and the onset of the following vowel, $F1$ was estimated to transition from 0 to 200 Hz (corrected, 180 to 380 Hz) and $F2$ from 1600 to 1900 Hz as the lip aperture increased from 1.4 to 2.5 cm^2. The utterances detour and veto were examined qualitatively for evidence of coarticulation. Anticipation of the rounded lip position required for V_2 was found to influence not only the lip movements of /t/, but also, to some extent, the lip position of the preceding vowel, /i/.

Fujimura also performed a second study, once again using the stroboscopic motion picture system (Fujimura, 1961b). The study investigated the production of bilabial stop and nasal consonants in a number of English words, each containing the consonant /p/, /b/ or /m/ followed by the vowel /i/, /o/ or /ɑ/, such as speech, mope and bock. Data from a single speaker were filmed at an exposure rate of 240 Hz. An important outcome of the study was the identification of a three-stage transition in the articulator movements following the stop release, from which a corresponding three-stage transition in the first formant frequency $F1$ was inferred. The first stage occurs during the 5 - 10 msec immediately after the lips begin to open. It consists of an abrupt increase in lip opening cross-sectional area, at a rate of approximately 100 cm^2/sec. Fujimura proposed a model that predicted a corresponding rapid upward shift in $F1$. Within just the first 5 msec, the model predicts $F1$ will shift from 0 Hz (assuming the vocal-tract walls are rigid) to a value of 200 - 400 Hz, depending upon the following vowel. (The model is discussed in some detail in Chapter 2, Section 2.2.) The second stage consists of a slower increase in the lip opening cross-sectional area, rate approximately 25 cm^2/sec, corresponding to a slower rise in $F1$. Fujimura observed that, as the air pressure in the mouth is released at the time of lip opening for the stop, a highly damped oscillation of the lips occurs with a frequency of 35 - 40 Hz. The slower transition in the second stage is partly attributable to the lips making their first inward vibratory movement. During the second stage, the rate of movement of the lips was found to be comparable to the rate of downward movement of the jaw. In general, a third stage is present for stops, where the lip separation is again relatively rapid, although not as rapid as in the initial stage. Movement of the jaw may also contribute to the $F1$ transition in this final stage. The rate of the $F1$ shift is markedly lower in the third stage than that of the first stage. It should be noted that, in contrast to the influence of the lip and jaw movements on $F1$ throughout all three stages, the first formant is not very sensitive to changes in the shape of the tongue, such as tongue movements made in anticipation of the following vowel. Variations between bilabial and nasal consonant productions, as well as effects of phonetic context, were observed to influence the three transition stages. Stops in the word-initial position exhibited the three stages most distinctly. Intervocalic stops also showed three separate stages, with the absolute speed for the first transition not as great as that of the initial stop. The nasal consonant /m/, with no pressure buildup, does not demonstrate the lip vibrations seen in the production of the stops /p, b/. For /m/, the transition during the first stage is not as rapid as for the initial stop; however, the speed in the second stage is considerably greater than that of the second stage for stops, due to the lack of the vibratory component. The second stage in the production of the nasal /m/ is analogous to the third stage in stop production. As a result, nasal production typically does not exhibit a third stage. For initial and intervocalic /m/, the average rate of change of area during the first 20 msec, including the first stage, is typically equal to or slightly greater than that for initial /p/.

In each of the stroboscopic motion picture studies, recording speech movements via the use of a videocamera is limited to recording only the visible articulators, such as the lips and jaw. In the first study, information about the primary articulators was not attainable, and in both studies the lengths of the constrictions were not determined. In contrast, the lip constriction cross-sectional area was readily obtained; however, the lip opening cross-sectional area recorded by a videocamera will not necessarily be accurate if the plane of the lip opening is not kept perpendicular to the line from the camera. A source of possible error in the jaw displacement measurements is the ability of the facial muscles in the chin to move independently of the jaw. Thus, whenever the lower incisors could be seen between the lips, their movement was used as a more accurate reflection of lower jaw position. The problems inherent to the use of a low frame rate were present in the first study, but not the second, which used a faster frame rate. The stroboscopic motion picture technique reduced the blurring in the spatial domain via the use of a strobe light to effectively shorten the frame exposure time. In addition, the system has the advantage of being completely noninvasive.

Williams (1995) utilized acoustics, in the form of a perceptual judgment test, to determine listener-preferred rates of constriction cross-sectional area change with time following voiced stop-consonant release. Syllables of the form /Cɑ/, with C being one of /b, d, g/, were synthesized using a set of articulatory parameters to control the increase in constriction cross-sectional area after release. The selection of the particular rates of area increase was based on the above-mentioned studies by Fujimura (1961b), Kent and Moll (1972), and Kuehn and Moll (1976), in conjunction with Stevens (Acoustic Phonetics, in preparation). Fujimura determined a rate of change of cross-sectional area for a labial stop of 100 cm^2/sec. Based on the work of Kent and Moll (1972), Kuehn and Moll (1976), and from acoustic data, Stevens derived rates of 50 - 100 cm^2/sec for an apical stop or nasal consonant and 25 cm^2/sec for a velar stop consonant. In the synthesis experiment performed by Williams, a set of five different rates of increase (10, 15, 25, 50 and 100 cm^2/sec) was chosen for the labial and alveolar voiced stops and a second, slower set (10, 15, 20, 30 and 50 cm^2/sec) for the velar stop. The listeners indicated a strong preference for the faster rates of increase (25, 50 and 100 cm^2/sec) for the labial stop. In the case of the alveolar stop, the preference was strongly for the 25 cm^2/sec rate. The velar stop results were a little less clear, but preference seemed to be for the 30 cm^2/sec rate. Overall, the results showed that intermediate rates of increase were preferred for all places of articulation (faster rates were also preferred for labials). Slower rates were generally not preferred. The conclusion of the study was that it is possible that a single constriction cross-sectional area rate of increase following stop release (about 30 - 40 cm^2/sec) could be used to synthesize all voiced stops.

1.3 Thesis Outline

The literature survey reveals a notable lack of detailed acoustic analysis of stop-consonant production. The objective of the present research is to utilize analysis of experimental acoustic data, supplemented by measurements of articulator movements, to refine and expand existing theoretical models of stop production. Whenever possible, attempts will be made to quantify various aspects of the relationship between the articulator movements and the resultant acoustics. Stop-consonant production will be investigated using three techniques: theoretical modeling (Chapter 2), articulation experiments (Chapter 3) and acoustic analysis (Chapter 4).

In Chapter 2, two theoretical models will be introduced. The first is a low-frequency circuit model from which the noise burst sound source can be predicted with the aid of an additional equation. The second is a high-frequency tube model of the filtering function of the vocal tract. Model parameters will be constrained based on anatomical, physiological and acoustic data, including measurements of articulatory movements, as obtained in Chapter 3, and measurements of acoustic data as obtained in Chapter 4.

In Chapter 3, the technique used to obtain the experimental measurements of the articulator movements via a device called an electro-magnetic midsagittal articulometer (EMMA) will be described. The data processing procedures performed on the articulation data will be presented. The articulator displacements and velocities will be analyzed to determine certain aspects of how stop consonants are produced. The effects of phonetic context on stop-consonant production will also be discussed.

In Chapter 4, analysis will be performed on the acoustic data, which were recorded simultaneously with the EMMA articulatory data. The low-frequency circuit model presented in Chapter 2 will incorporate the articulation data, in the form of a parameter estimating the linear rate of change in the time-varying constriction cross-sectional area following stop-consonant release. The estimate of the linear rate of area increase will be fine-tuned via matching the duration of the modeled noise burst to the noise burst duration measured from the acoustic data. The first formant frequency ($F1$) transition will be calculated using the linear area estimate as an input parameter in a program which solves the wave equation for the high-frequency model introduced in Chapter 2. The calculated $F1$ will be used to supplement the measured $F1$ transition during the time period immediately following the release, when acoustic information is limited.

In Chapter 5, a summary of the results from Chapters 2, 3 and 4 will be presented, and directions for future research will be outlined. Implications for speech recognition, synthesis and studies of disordered speech production will be identified.

Chapter 2

Theoretical Considerations

Theoretical models have been developed to describe the aerodynamic, acoustic, and mechanical events occurring during the three phases of stop-consonant production. The models can be classified according to the frequency ranges involved. The low-frequency model accounts for the vocal-tract pressures and airflows generated by the relatively slow-moving articulators. Based upon knowledge of the pressures and flows, the noise source created upon release of the constriction can also be estimated. The high-frequency model accounts for the filtering of the acoustic signal by the vocal tract and the resultant acoustics produced.

2.1 Model of the Frication Noise Source

A theoretical, low-frequency model which examines vocal-tract movements, airflows and pressures occurring during the three phases of stop-consonant production was proposed by Stevens (1993). Based on physiological information about the vocal tract and knowledge of the articulator movements, the model predicts the average pressures and airflows generated in the vocal tract throughout stop production. Stevens (1993) determined that the pressures and flows within the vocal tract can be estimated by modeling the vocal tract during consonant production as a tube with two constrictions, one at the glottis and one formed by an articulator within the vocal tract, as shown in Figure 2-1(a). A corresponding circuit diagram of the system is given in

Figure 2-1(b). The circuit model is similar to those developed by Rothenberg (1968), Westbury (1979), and Müller and Brown (1980).

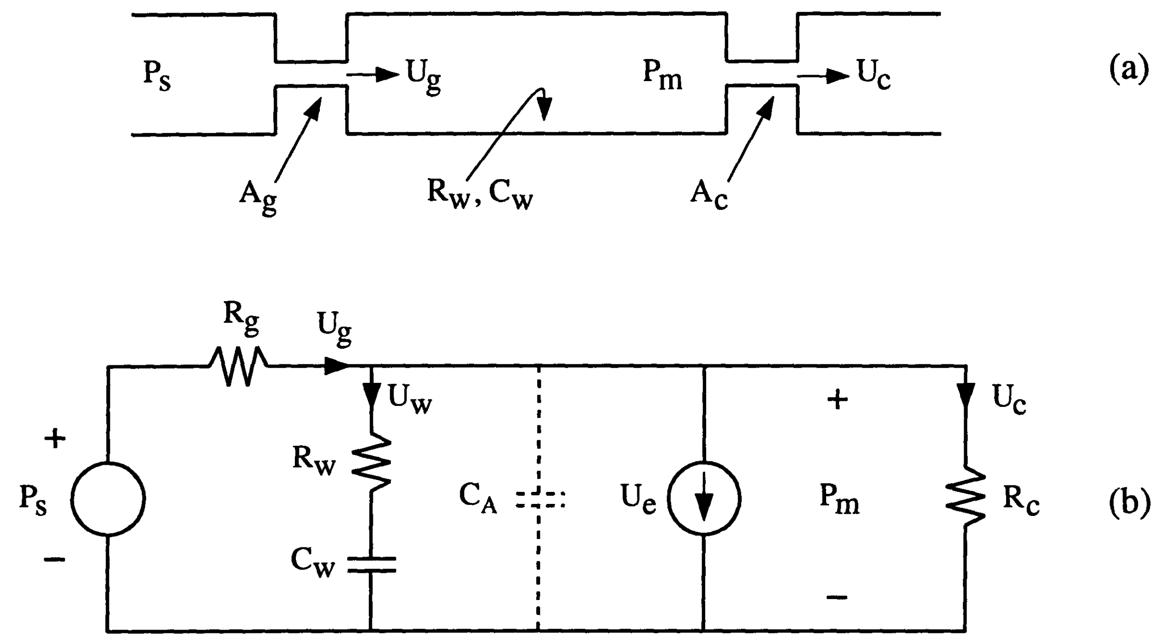

Figure 2-1: (a) Model for estimating average airflows and pressures during consonant production. (b) Equivalent circuit model. (Adapted from Stevens, 1993, Figure 2).

The variables shown in Figure 2-1 are defined:

P_s = Subglottal pressure source

R_g = Glottal resistance

A_g = Cross-sectional area of the glottis

U_g = Glottal airflow

R_w = Resistance of the vocal-tract walls

C_w = Acoustic compliance of the vocal-tract walls

U_w = Airflow due to inward and outward passive movement of the vocal-tract walls

C_A = Acoustic compliance of the vocal-tract air volume

U_e = Volume velocity source for active muscular contraction and expansion of vocal-tract walls

P_m = Pressure in the mouth

R_c = Constriction resistance

A_c = Cross-sectional area of the constriction

U_c = Airflow through the constriction

The subglottal pressure, P_s, is the principal driving source for the flow. P_s is assumed to be constant over the time interval of production of a stop consonant, and is estimated to be 8 cm H_2O. The assumption of a constant P_s is reasonable (within about 10%) as long as the airflow from the lungs does not exceed about 400 cm^3/sec, which is the case for normal speech production (Rothenberg, 1968). The vocal-tract walls are nonrigid, moving in response to changes in pressure within the vocal tract. At low frequencies (up to 30 - 40 Hz), the impedance of the walls can be approximated by an acoustic resistance R_w in series with an acoustic compliance C_w. Average values are estimated to be R_w = 10 dyne-sec/cm^5 and $C_w = 10^{-3}$ cm^5/dyne for labial and alveolar stop consonants, in which the total surface area of the vocal-tract walls posterior to the incisors is approximately 100 cm^2. For velar stops, average values are R_w = 15 dyne-sec/cm^5 and C_w = 8 x 10^{-4} cm^5/dyne, where the wall surface area is believed to be closer to 70 cm^2 (Ishizaka, French and Flanagan, 1975; Glass, 1986). (The effects of the nonrigid vocal-tract walls at higher frequencies are discussed in Section 2.2.) The acoustic compliance, C_A, is estimated to be 4 x 10^{-5} cm^5/dyne, one to two orders of magnitude smaller than C_w, so the effect of C_A is considered to be negligible. A speaker may actively expand or contract the volume of the vocal tract between the two constrictions during the production of some sounds. The effect of this expansion is represented by the volume velocity source U_e, which is positive if there is an active expansion and negative if there is an active contraction of the volume. It is assumed for the purposes of this study that there is a negligible amount of active expansion or contraction of the vocal-tract walls, so U_e is neglected. The acoustic resistances R_g and R_c are nonlinear, and, over much of the range of constriction sizes, can be approximated by a dynamic resistance proportional to volume velocity (Stevens, 1971). Expressions for the resistances appear in Equations 2.1 and 2.2.

$$R_g = \frac{\rho U_g}{2A_g^2} \tag{2.1}$$

$$R_c = \frac{\rho U_c}{2A_c^2} \tag{2.2}$$

When the constriction is extremely narrow, a second term must be added to each of the acoustic resistances R_g and R_c to account for the effects of the viscosity of the air. The viscous resistive component is non-negligible only within the first millisecond or so following the stop release, when the constriction is still very narrow. Since this thesis does not contain an in-depth analysis of the pressure and flow events occurring within the first millisecond following the release, the viscous component of the resistance will be neglected.

The thesis focuses in detail on the production of the labial and alveolar voiceless stop consonants in various phonetic environments. In-depth examination is performed of stop production during the time period between the release and the onset of the following vowel, in which limited acoustic information is available. The modeling approach is presented below for the labial, voiceless, unaspirated stop /p/, as in spot, as an example of the type of modeling which will be performed in Chapter 4 with the aid of the experimental data. A brief discussion follows, at the end of this section, of the relatively minor ways in which this modeling approach will vary to accommodate other phonetic environments for /p/, as well as the alveolar stop /t/ in various phonetic environments.

Example: Production of the Labial, Voiceless, Unaspirated Stop /p/

Production of the labial stop /p/, from the time of the stop release onward, can be modeled by the simplified circuit shown in Figure 2-2. Values have already been given for each of the circuit elements with the exception of the constriction cross-sectional areas, A_g and A_c (appearing in the formulas for R_g and R_c, respectively), which are time-varying and depend upon the stop produced, as well as its phonetic environment. Typically, the average area of the glottal constriction, A_g, changes with time in the

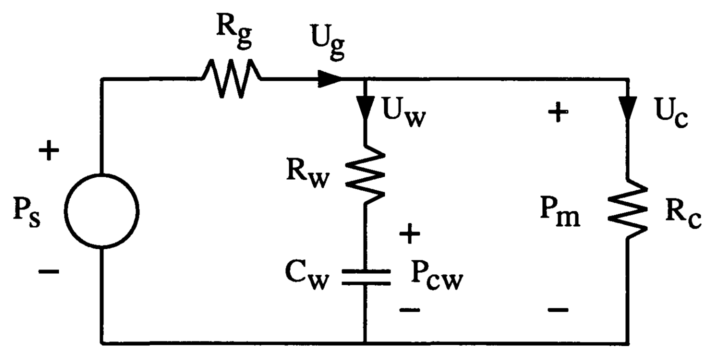

Figure 2-2: Circuit model for the production of /p/ following stop release.

vicinity of the times when the supraglottal constriction is formed or released. When pressure builds up above the glottis during closure for a consonant, outward forces are exerted on the vocal-fold surfaces, causing a passive increase in the glottal area. Upon consonant release, the intraoral pressure diminishes rapidly, resulting in a passive decrease in the glottal area since the outward forces holding the vocal folds open are no longer present. In addition, it is possible to have active adjustment of the glottal configuration during the production of the consonant, such as the adjustment required to maintain vocal-fold vibrations for a voiced stop. The production of an unaspirated /p/ requires the vocal folds to be slightly spread initially to prevent vocal-fold vibrations, yet not spread enough to produce significant aspiration. The normal range for the glottal area during aspiration or breathy voicing is 0.1 - 0.4 cm^2 (Stevens (Acoustic Phonetics, in preparation)). For the unaspirated stop consonants, an initial value of 0.1 cm^2 was chosen for A_g from the lower end of the range. During the time period between the release and the onset of the following vowel, the vocal folds are believed to be gradually moving closer together, due to a combination of a passive decrease in glottal area coupled with preparation for the vocal-fold vibrations of the upcoming vowel. The resulting A_g is assumed to be linearly decreasing from 0.1 to 0.05 cm^2 during this time period. The rate of decrease may vary, based on phonetic context. For the unaspirated stop consonants, A_g was set to decrease from 0.1 to 0.05 cm^2 during the first 40 msec following stop-consonant release, then to remain constant at 0.05 cm^2 thereafter.

An initial linear rate of area increase, A_c, will be estimated for the supraglottal constriction cross-sectional area from experimental articulation data, with the aid of models of the constriction opening. The initial linear rate will be refined based on output from the low-frequency circuit model and the frication noise source estimate introduced in this section, as well as experimental acoustic data. The details of the area derivation will be described fully in Chapter 4, Section 4.4, and will be outlined later in the present section for the example of the unaspirated stop /p/. The resultant linear area is approximately 47 cm^2/sec for the speaker in this example, as shown in Figure 2-3(a).

A set of equations based on conservation of mass and energy may be written for the simplified circuit of Figure 2-2. Kirchhoff's Current Law, the law of conservation of mass, may be applied to obtain Equation 2.3. Kirchhoff's Voltage Law, the law of conservation of energy, provides Equations 2.4 and 2.5. Equation 2.6 represents the constitutive relationship for the acoustic compliance.

$$\frac{P_s - P_m}{R_g} = \frac{P_m - P_{cw}}{R_w} + \frac{P_m}{R_c} \tag{2.3}$$

$$P_m = U_w R_w + P_{cw} \tag{2.4}$$

$$P_m = -U_g R_g + P_s \tag{2.5}$$

$$U_w = C_w \frac{dP_{cw}}{dt} \tag{2.6}$$

P_{cw} is the pressure drop across the acoustic compliance C_w. The capacitor representing the compliance C_w is assumed to be fully charged by the time of the stop release, yielding an initial intraoral pressure P_m of 8 cm H_2O.

It is possible to obtain closed-form solutions for the average pressures and airflows of the vocal tract. Arbitrarily, the choice was made to express each of the pressures and flows in terms of U_w. With this selection, the average pressures P_m and P_{cw}

in Equations 2.4 and 2.6 are already expressed in terms of only U_w and constants. Combining Equations 2.1, 2.2, 2.4, and 2.5 results in Equations 2.7 and 2.8, expressing each of the airflows U_g and U_c in terms of U_w, the area functions A_g and A_c, and constants. Each of A_g and A_c are known functions of time a priori for the entire interval following the release.

$$U_g = \sqrt{\left(\frac{2A_g^2}{\rho}\right)(P_s - U_w R_w - P_{cw})} \tag{2.7}$$

$$U_c = \sqrt{\left(\frac{2A_c^2}{\rho}\right)(U_w R_w + P_{cw})} \tag{2.8}$$

To find the closed form solution for U_w, each of Equations 2.4, 2.6, 2.7, and 2.8 are substituted into Equation 2.3, yielding a nonlinear differential equation in terms of only U_w, the area functions A_g and A_c, and constants. The differential equation can be solved for U_w, which will vary with time because each of the area functions A_g and A_c vary with time. Once the airflow U_w is known, Equations 2.7 and 2.8 can be solved to yield the airflows U_g and U_c, respectively. The pressures P_m and P_{cw} can be found from any combination of two of Equations 2.4, 2.5 and 2.6. In this thesis, the average pressures and airflows in Equations 2.3 - 2.6 were obtained by finding solutions iteratively via a computer program.

The pressure drop across the supraglottal constriction, P_m, and the airflow through the supraglottal constriction, U_c, for the time period following the /p/ stop release are shown in Figure 2-3 (b) and (c), respectively. Upon release, the model predicts a rapidly decreasing intraoral pressure and an airflow which increases from zero to a peak value quickly, then decreases to reach steady state by the time of the following vowel. The airflow U_c consists of two components: U_w due to the rapid inward movement of the walls after the pressure is released, and the glottal flow U_g (Figure 2-3(c)). For the production of the unaspirated /p/, with a linear rate of area increase of about 47 cm^2/sec in this example, U_w contributes more than U_g to both the amplitude and shape of U_c during the first 25 msec or so following the stop release.

As a consequence of the changing airflows and pressures following release, various

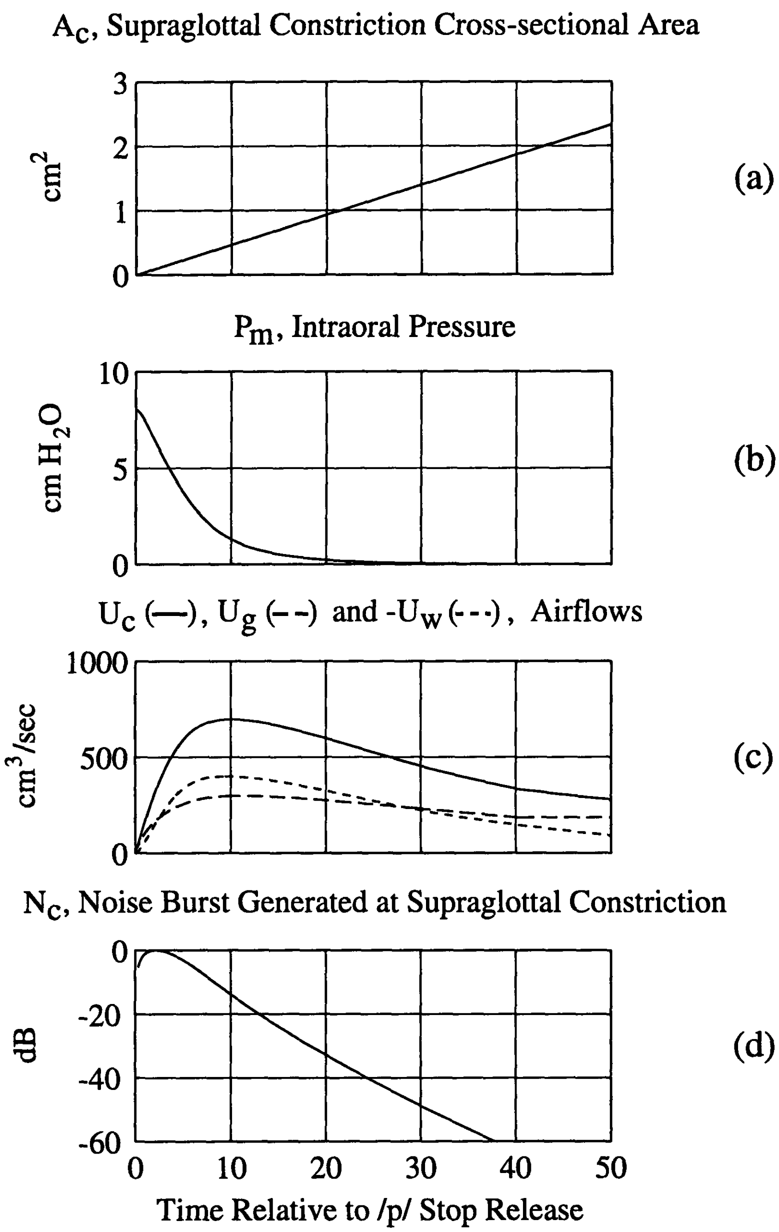

Figure 2-3: The unaspirated stop consonant /p/ upon release of closure: (a) Lip-opening constriction cross-sectional area, A_c (as determined in Chapter 4); (b) Pressure within the mouth, P_m; (c) Airflow through the lip-opening constriction, U_c (solid line), airflow through the glottis, U_g (dashed line), and airflow generated by the inward displacement of the vocal-tract walls, -U_w (dotted line) (the negative sign indicates the direction of displacement of U_w is inward); (d) Frication noise burst, N_c. Time zero is the instant of stop release. The stop /p/ was spoken in the utterance "Say spot again." by Subject 1.

types of noise sources are generated in the vocal tract. During stop-consonant production, pressure builds up behind a closure at a supraglottal location in the vocal tract, then the closure is released. The current theoretical model (Stevens, 1993) proposes the existence of a sequence of three different types of radiated noise following release of the pressure buildup. The first is the transient sound as the compressed air in the vocal tract is expelled, the second is the frication noise burst generated at the supraglottal constriction, and third is the aspiration noise which arises from turbulence near the glottis, causing transitions to become apparent in the formants. A schematic representation of the three types of sound sources following release appears in Figure 2-4.

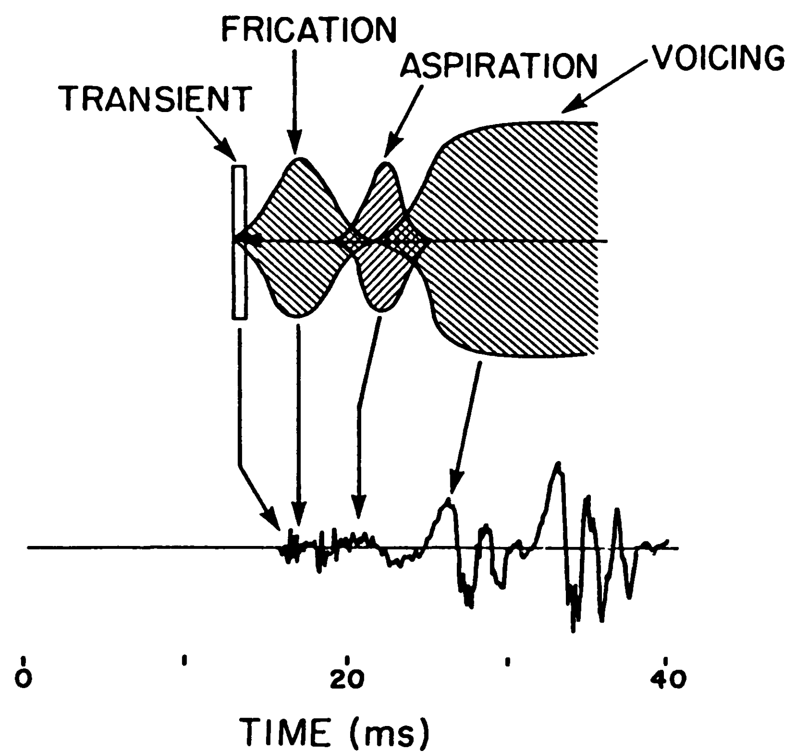

Figure 2-4: Schematic representation of sequence of events at the release of a voiceless unaspirated stop consonant. A typical waveform (with time scale) is shown at the bottom. Reprinted with permission from Stevens (1993).

During the initial 5 - 10 msec following the release, rapid airflow through the narrow supraglottal constriction generates turbulence, creating a frication noise source.

Stevens (1993) proposed a model for estimating the frication noise source, based on the work of Fant (1960), Stevens (1971), Shadle (1985), Pastel (1987), and others. In the model, the turbulence noise source is represented as a sound-pressure source near an obstacle downstream from the constriction. Based on knowledge of the constriction cross-sectional area, A_c, and either the airflow through the constriction, U_c, or the pressure drop across the constriction, P_m, following the release (refer to Figure 2-3 (a) - (c)), the amplitude of the noise source can be calculated. The source amplitude is approximately proportional to $U_c^3 A_c^{-2.5}$, based on empirical data with some theoretical backing. (The formula can also be expressed in terms of pressure, $P_m^{1.5} A_c^{0.5}$, with the aid of Equation 2.2 and Ohm's Law.) The model output of the frication noise source appears in Figure 2-3(d). The noise source rises to a peak within 1 - 2 msec following the release, then decreases rapidly as the cross-sectional area of the constriction increases.

The supraglottal constriction cross-sectional area derived from the experimental articulation data, as described in Chapter 4, Section 4.4, will be refined with the aid of the models introduced in this section and the experimental acoustic data in the following manner. First, an estimated linear rate of increase for the cross-sectional area is utilized as a parameter in the low-frequency circuit model. The circuit model output and the area are then inputs to the model of the frication noise source. The duration of the modeled noise burst is compared to the duration of the noise burst measured from the experimental acoustic data (refer to Chapter 4, Section 4.3). The constriction cross-sectional area is adjusted, via a constant multiplier, until the corresponding linear rate of area increase produces a modeled noise burst equal in duration to the measured noise burst (to the nearest millisecond). For the unaspirated /p/, the resultant supraglottal constriction cross-sectional area and its corresponding linear rate of area increase, A_c, are shown in Figure 2-5. (The linear area is also shown in Figure 2-3, in association with the model outputs produced by that particular value of area increase.)

This example illustrates the process of modeling the production of the frication noise burst which follows an unaspirated /p/. The example also demonstrates the application of the models to the refinement of the linear rate of constriction cross-sectional area increase, A_c. The procedures shown in this example will be followed for /p/ and /t/ in various phonetic environments, in Chapter 4. Values of A_g and A_c may vary, depending upon the stop consonant and its phonetic environment.

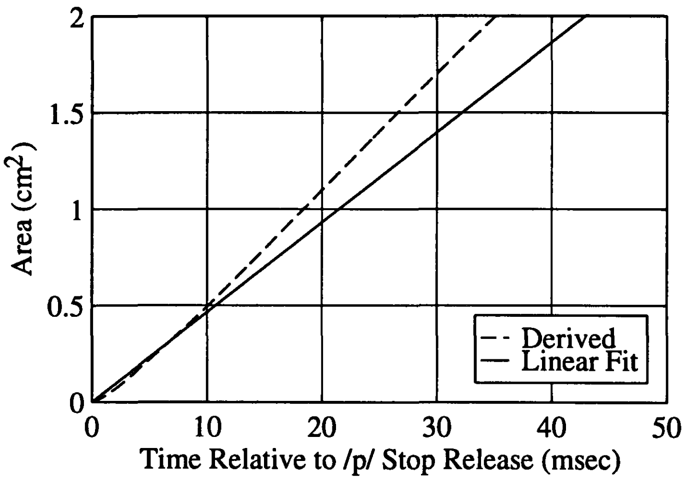

Figure 2-5: Labial constriction cross-sectional area for the utterance spot, based on data derived from Subject 1. The dashed line is the area derived from the articulation data, models, and acoustic data, as discussed in the text. The solid line is the best linear fit to the first 10 msec of the derived area. The linear rate of area increase is 39.0 cm^2/sec.

2.2 Model of the Vocal-Tract Filter

A theoretical, high-frequency model of the vocal tract has been developed to predict the resonant frequency transitions occurring in the acoustic signal following the stop-consonant release. The continually-changing shapes assumed by the vocal tract during speech production are modeled by sets of tubes with varying cross-sectional areas, as in Figure 2-1(a) for consonant production. Sound sources, whether noise or glottal-pulse generated, excite the resonant frequencies of the portion of the vocal tract downstream from the source. The downstream vocal-tract cavity acts as a continuously-changing filter, filtering the sound passing through that section of the vocal tract. The process of source generation coupled with vocal-tract filtering produces the resultant acoustic signal. The high-frequency model is valid for a frequency range of approximately 150 - 6000 Hz. Predictions can be made, based on the vocal-tract filter model, as to when and at what rates the resonant, or formant, frequencies in the resultant acoustics will change with time following the stop-consonant release.

A simplified version of the high-frequency vocal-tract filter model was first proposed by Fujimura (1961b). The model addresses the time period immediately after the release of a labial stop consonant. The model represents the articulator movements during production of the initial portion of the frication noise burst following the release, from which the timing and rate of the corresponding first formant frequency ($F1$) transition can be inferred. For the first few milliseconds following the labial stop release, Fujimura observed an abrupt initial increase in lip opening cross-sectional area. The changing shape of the vocal tract immediately following the labial stop release was modeled by Fujimura as a Helmholtz resonator (Figure 2-6). In Figure 2-6,

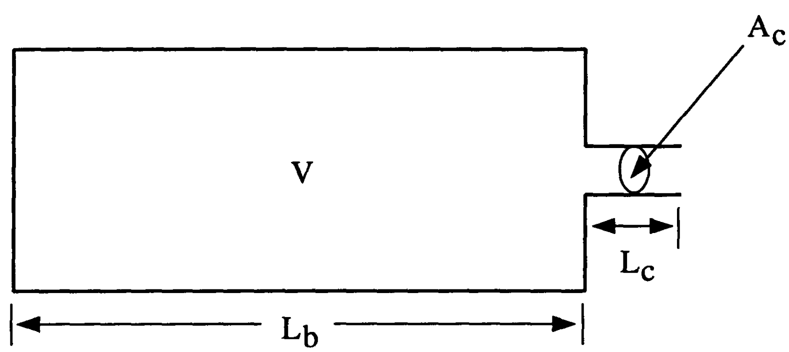

Figure 2-6: Helmholtz resonator vocal-tract filter model for a labial stop, valid for first few milliseconds following stop release. L_b is the length of the back cavity, typically about 17 cm for a male speaker and L_c is the length of the constriction, approximately 0.5 cm for /p, b/. V is the volume of the back cavity, approximately 70 cm^2 for a male, and A_c is the lip opening cross-sectional area, which varies with time.

the glottis is on the left, modeled as closed, and the lips are on the right, represented by the short, front tube. The lower lip movement was found by Fujimura to be fairly independent of the lower jaw movement during this time interval. The narrow front tube of the Helmholtz resonator, therefore, represents movement of both the lips and the lower jaw, since the recorded lip movement during this time period consists of the superposition of the lip movement independent of the jaw and the lower lip movement due to the lower lip riding on the lower jaw. Both the length of the constriction and the volume of the back cavity, the cavity behind or upstream from the constriction, are assumed to be constant. The constriction is modeled as an acoustic mass and the back cavity as an acoustic compliance. The simple, lumped-element resonator is only valid within approximately the 150 - 300 Hz range, consequently it provides a first approximation of the $F1$ transition during only the initial few milliseconds following release. From circuit theory, the equation corresponding to the movement of the first formant frequency in the Helmholtz resonator model is Equation 2.9. This equation assumes nonrigid walls with some mass. The first formant frequency $F1$ can be calculated from the equation if the changing cross-sectional area A_c is known.

$$F1 = \sqrt{(F1_c)^2 + \left(\frac{c}{2\pi\sqrt{\frac{VL_c}{A_c}}}\right)^2} \tag{2.9}$$

The $F1_c$ term in Equation 2.9 represents the effect of the nonrigid vocal-tract walls on the value of $F1$ during the time period immediately following the stop release. The vocal-tract walls, consisting of the surfaces of the tongue, cheeks and pharynx, have some mass and resistance, and therefore have a finite, nonzero impedance. The specific acoustic impedance (acoustic impedance per unit area) is given by Equation 2.10, where R_{sw} and X_{sw} are both large compared with ρc, the specific acoustic impedance of air. (Stevens (Acoustic Phonetics, in preparation); Ishizaka et al., 1975; and others). (It should be noted that $R_{sw} = R_w \times A_w$, where R_w is the acoustic resistance introduced in Section 2.1.)

$$Z_{sw} = R_{sw} + jX_{sw} \tag{2.10}$$

Estimates of the impedance of the inner surfaces of the vocal tract have been made from data on the impedance of the skin on other parts of the body. In the frequency range of the first formant for vowels, this impedance (per unit area) is approximately $Z_{sw} = 1000 + j2\pi f \times 2.0$ dynes $\cdot$ sec/cm^3 (Ishizaka et al., 1975; Wodicka et al., 1993). When the vocal-tract configuration is constricted, yielding a low $F1$, the mass reactance of the walls can cause a significant shift in $F1$ (Fant, 1972). The amount of the shift is greatest for a completely closed vocal tract, as in the configuration for a stop consonant during the closure interval. A circuit model representing this configuration for a labial stop consonant appears in Figure 2-7. In this circuit model, the wall impedance is represented by an acoustic resistance in series with an acoustic mass, appropriate for the frequency range 100 - 200 Hz discussed in this section. In contrast, the wall impedance was represented by an acoustic resistance in series with an acoustic compliance when the frequency range under discussion was 30 - 40 Hz, in Section 2.1. The natural frequency $F1_c$ of the circuit in Figure 2-7 is

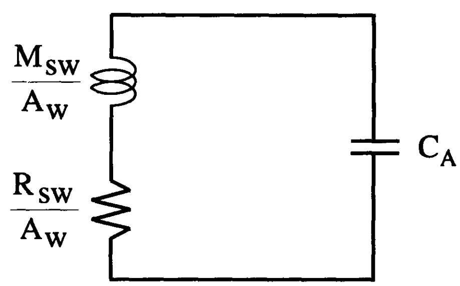

Figure 2-7: Low-frequency circuit model of vocal tract with closed lips and glottis. A_w is the surface area of the vocal-tract walls, M_{sw} and R_{sw} are mass and resistance, respectively, of walls per unit area, and C_A is the acoustic compliance of the vocal-tract volume.

determined by the mass M_{sw}, the area A_w and the acoustic compliance C_A of the closed cavity (Equation 2.11). For a vocal-tract surface area A_w of 90 cm^2, mass M_{sw} of 2 gm/cm^2, and a vocal-tract volume of 45 cm^3 yielding an acoustic compliance C_A of 3 x $10^{-5} cm^5/dyne$, $F1_c$ is calculated to be approximately 190 Hz.

$$F1_c = \frac{\sqrt{A_w}}{2\pi\sqrt{M_{sw}C_A}} \tag{2.11}$$

Actual measurements of the resonant frequency for a bilabial stop configuration give values of about 180 Hz for an adult male talker and about 190 Hz for a female talker (Fujimura and Lindqvist, 1971; Fant et al., 1976). The value of $F1_c$ would be zero if the walls had infinite impedance, thus the influence of nonrigid walls on $F1$ during the closure interval for a stop consonant is considerable, increasing the value of $F1$ for a closed volume from 0 Hz to approximately 180 Hz, for a male speaker.

An additional factor should be added to the value of $F1$ in Equation 2.9 when /p/ is aspirated. During the production of an aspirated stop, the vocal folds are held apart and are not accurately modeled by a closed glottis. The relatively open glottis no longer has infinite impedance. The finite glottal impedance has resistive and reactive components. The reactive component causes an upward shift in the formant frequencies of the vocal tract. The relative shift is greatest for $F1$. The amount of shift depends to a certain extent on the degree of aspiration. If the vocal tract is represented by a uniform tube of length 17.7 cm, cross-sectional area 3 cm^2 and maximum glottal area 0.11 cm^2 (appropriate for the relatively unaspirated /p/ in spot), the value of $F1$ shifts upward by 30 Hz, neglecting the effect of the subglottal impedance (Stevens (Acoustic Phonetics, in preparation)). The shift would be even greater for the more aspirated /p/ in pot.

A model similar to the model developed by Fujimura for the production of a labial stop could be proposed for the vocal-tract shape of an alveolar stop during the time period immediately following the release. This vocal-tract filter model is shown in Figure 2-8. The effect of the short front tube can be neglected in the determination of $F1$ for the first few milliseconds following the release, since the constriction cross-sectional area, A_c, is very small compared to the cross-sectional area of the front tube during that time interval. Consequently, the initial $F1$ transition can be determined using Equation 2.9, the same equation used to determine the initial $F1$ transition for the labial stop consonants.

Although the Helmholtz model represents a simplified view of the changing vocal-tract shape immediately following a labial or alveolar stop-consonant release, it is presented in the thesis because it lends some insight into how vocal-tract configuration changes result in shifts in the first formant frequency. In order to determine the first formant frequency transition for all time following the stop release, however, the wave equation must be solved. The following description of the derivation of the wave equation is adapted from a discussion in Stevens (Acoustic Phonetics, in prepa-

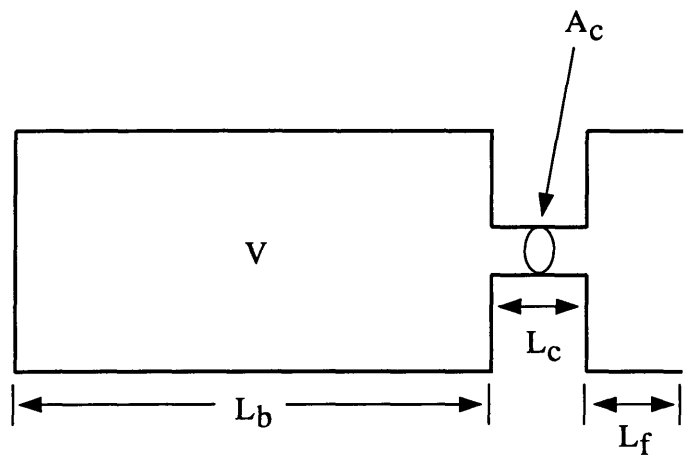

Figure 2-8: Vocal-tract tube filter model for an alveolar stop. L_b is the length of the back cavity, typically about 15 cm for a male speaker, L_c is the length of the constriction, about 0.5 cm for /t, d/, and L_f is the length of the front cavity, typically about 2 cm. V is the volume of the back cavity, approximately 60 - 70 cm^2 for a male, and A_c is the time-varying cross-sectional area of the constriction formed by the tongue tip and the hard palate.

ration). The sound pressure $p(x,t)$ and volume velocity $U(x,t)$ for one-dimensional wave propagation in an acoustic tube are related by Equations 2.12 and 2.13, derived from Newton's law and compressibility considerations (cf. Beranek, 1954).

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A}\frac{\partial U}{\partial t} \tag{2.12}$$

$$\frac{\partial U}{\partial x} = -\frac{A}{\gamma P_o}\frac{\partial p}{\partial t} \tag{2.13}$$

The variable A is the cross-sectional area of the tube at the point x, P_o is the ambient pressure, ρ is the ambient density of the air (0.00114 gm/cm^3 for air at body temperature), and γ is the ratio of specific heat at constant pressure to specific heat at constant volume ($\gamma = 1.4$ for air). If exponential time dependence is assumed for the pressure and volume velocity in Equations 2.12 and 2.13, and U is eliminated from each equation, the result is Equation 2.14, referred to as the Webster horn equation

(Morse, 1948), or the wave equation.

$$\frac{d^2p}{dx^2} + \frac{1}{A}\frac{dA}{dx}\frac{dp}{dx} + k^2p = 0 \tag{2.14}$$

The velocity of sound $c = \sqrt{\frac{\gamma P_0}{\rho}}$ and the wave number $k = \frac{2\pi f}{c}$. (For air at body temperature, c = 35,400 cm/sec.) Solutions to the wave equation can be found in closed form for only a few area functions $A(x)$, where $A(x)$ is the cross-sectional area along the entire length of the vocal tract.

The natural frequencies of an acoustic tube with area function $A(x)$ are the values of the frequency f for which Equation 2.14 has a solution, subject to the boundary conditions at either end of the tube. S. Maeda developed a software program for an IBM PC-compatible computer which solves the wave equation for an arbitrary area function $A(x)$ of the vocal tract. The program partitions the vocal tract into many short, juxtaposed tubes of constant cross-sectional area, and the wave equation is solved for each short tube, subject to the boundary conditions at both ends of the short tube. The length of each short tube is arbitrary. For a given $A(x)$, the solution to the wave equation is assumed to be quasistatic, i.e. the rate of change of the vocal-tract shape is slow compared with the rate of change of the natural frequencies. When the program is given $A(x)$ for several consecutive instants in time following the stop release, the first five formant frequency transitions are generated. The program assumes the glottis remains closed. Several sources of loss in the vocal tract, including the radiation impedance at the mouth opening, the impedance of the vocal-tract walls, and viscous friction and heat conduction at the walls, are incorporated into the program.

Maeda's program will be utilized in Chapter 4 of the thesis to calculate the $F1$ transition following release for the voiceless, unaspirated labial and alveolar stop consonants. An estimate of the linear rate of constriction cross-sectional area increase with time following release will be derived from the experimental articulation data, as described in Section 2.1. Assumptions will be made about $A(x)$ for the remaining portions of the vocal tract, based on Fant (1960). Following the stop release, the first formant frequency is usually not evident in the acoustic waveform until the aspiration noise excites the vocal-tract resonances, typically 15 - 30 msec after release for the voiceless unaspirated stop consonants involved in this study. The $F1$ transition calculated by Maeda's program will be used to supplement the $F1$ transition measured from the acoustic signal. An example of the calculated and measured $F1$ transitions for the unaspirated /p/ in spot is shown in Figure 2-9. For comparison purposes, the $F1$ transition determined via the Helmholtz equation, Equation 2.9, is also included. (The 30 Hz upward shift in $F1$ due to the reactive component of the finite glottal impedance has been added to the values of $F1$ computed by Equation 2.9.) For about the first 2 - 3 milliseconds following the release, the Helmholtz equation represents the rate of the $F1$ transition fairly well; however, as time progresses, the vocal-tract shape is modeled less and less well by the Helmholtz resonator.

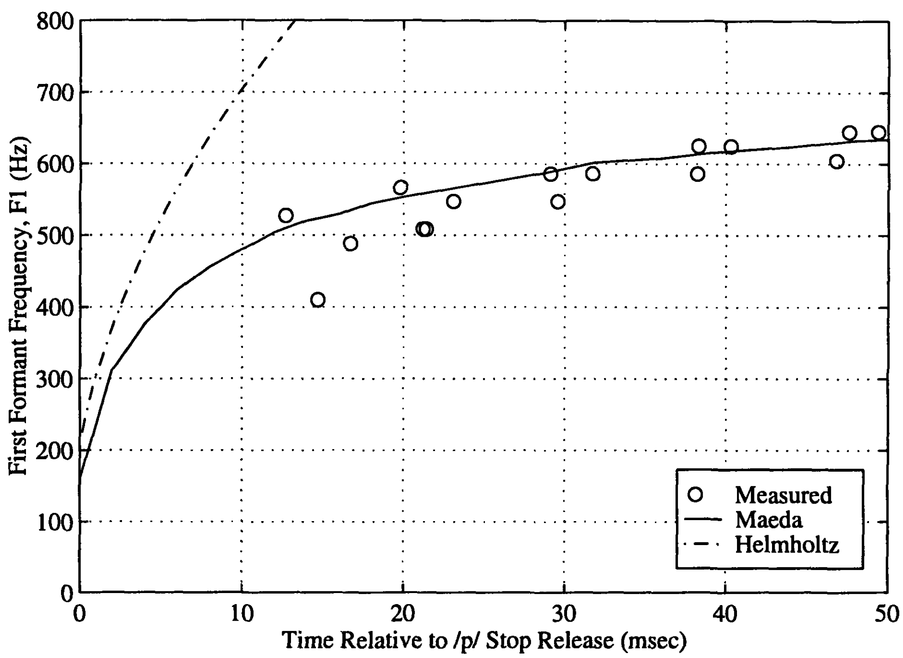

Figure 2-9: First formant frequency ($F1$) transition following release of the unaspirated /p/ in spot, spoken by Subject 1. Dots represent values of $F1$ measured from the acoustic data for all repetitions. The solid line is the estimate of the $F1$ transition calculated by Maeda's program, using the vocal-tract model of Figure 4-9 and the linear rate of area increase shown in Figure 2-3(a), derived from articulation and acoustic data. The dash-dot line is the $F1$ transition calculated from the Helmholtz equation.

2.3 Summary and Discussion

In Section 2.1, the process for determination of a linear rate of supraglottal constriction cross-sectional area increase with time, A_c, following the stop-consonant release was outlined. First, an initial estimate of the supraglottal constriction cross-sectional area and a corresponding linear rate of increase were derived from the experimental data (refer to Chapter 4, Section 4.4). The linear rate of area increase, A_c, was next utilized as a parameter in a low-frequency circuit model. An airflow output from the circuit model and A_c were then used as inputs to a second model, of the frication noise burst sound source. This model predicted that the noise burst amplitude would rise to a peak within a few milliseconds following the stop release, then fall off rapidly. The constriction cross-sectional area estimate, and its corresponding linear rate of area increase, were refined by comparing the duration of the modeled noise burst to that of the noise burst measured from the experimental acoustic data (refer to Section 4.3), and adjusting the constriction cross-sectional area, via a constant multiplier, until its corresponding linear rate of increase, A_c, resulted in the production of a modeled noise burst equal in duration to the measured noise burst, to the nearest millisecond.

In Section 2.2, a high-frequency model of the vocal-tract configuration following stop-consonant release utilizes as one of its parameters the linear rate of supraglottal constriction cross-sectional area increase, A_c, as determined in Section 2.1. The high-frequency model predicts the first formant frequency, $F1$, transition after release. The $F1$ transition calculated from the model is used to supplement $F1$ values measured from the experimental acoustic data.

The next chapter, Chapter 3, contains a description of how the experimental articulation data is measured and processed. A discussion of the effects of phonetic environment and inter- and intra-speaker variability on voiceless, unaspirated labial and alveolar stop consonant production is included. In Chapter 4, the linear rate of supraglottal constriction cross-sectional area increase is derived from the experimental data with the aid of models and experimental acoustic data. The linear constriction cross-sectional area rate of increase is used as a model input to the low- and high-frequency models, as discussed in the present chapter. The theoretical model outputs are then compared to various aspects of the experimentally-recorded acoustic signal.

Chapter 3

Articulation Analysis

The methods and results of the articulation analysis are presented in this chapter. The first section provides a description of the speakers' backgrounds and the corpus. The second section discusses the recording equipment, an electro-magnetic mid-sagittal articulometer, and method utilized in the study. In the third section, the data processing procedure developed to aid in evaluating the displacement data is described. The third section also examines the coordination, timing, sequence and rates of articulator movements. The fourth section focuses on the rates of the articulator movements in further detail, and the fifth section summarizes the results of the chapter.

3.1 Speakers and Corpus

Three male speakers participated in the study. They spoke American English as their first language, without any pronounced regional dialect. They had no known speech or hearing abnormalities. They were adult, in the age range approximately 18 - 45 years old.

The focus of the study is on the production of the labial and alveolar voiceless stop consonants. A corpus was created in which these two stops were examined in several different phonetic environments. The tokens consisted of single-syllable /CVt/ sequences which occur naturally in the English language. The syllable /CVt/

was composed of a voiceless stop consonant /p, t/, followed by a vowel /ɑ, i/, and terminated by the stop consonant /t/. The carrier phrase, "Say _____ again." was used to place the syllable in the middle of a phrase, thus avoiding the effects of the intonational rise or fall likely to occur at the beginning or end of a phrase. The terminal stop consonant /t/ was chosen to provide a distinct, consistent ending to each token within the carrier phrase. A sample utterance is, "Say pot again."

Two contrasting phonetic environments are provided by the vowels /ɑ, i/. To produce the vowel /ɑ/, the tongue body must be in a low, back position and the jaw in a low position. To produce the vowel /i/, the tongue body must be in a high, front position, and the jaw in a high position. A third phonetic environment is created by preceding half the tokens by the fricative /s/. During the production of /s/, the jaw position is fairly invariant across repetitions, contexts and speakers (Engstrand, 1980; Ichikawa et al., 1993). The fricative /s/ is included in the study to determine the effect that a severely-constrained environment preceding the stop has on stop production. In addition, the presence of /s/ prior to a stop causes the stop to be produced relatively unaspirated. The glottis is not open very wide at the time of the stop release, reducing the amount of aspiration noise prior to the vowel onset. Consequently, the aspiration noise burst for /p/ in <u>spot</u> is less pronounced than that for /p/ in <u>pot</u>. Other stops, as well as nasals, are included in the study to furnish a basis for comparison, including the velar stop /k/ and the labial nasal /m/. The voiced stops /b, d, g/ and the alveolar nasal /n/ were also recorded by the third subject. The voiceless stop /k/ was included so that the study could compare and contrast specific aspects of all the voiceless stops. The voiced stops were added in order to determine if there were measurable differences in articulator movements between voiceless and voiced stops. The nasal consonants were incorporated to compare the presence of pressure buildup in the mouth during stop production to the absence of similar pressure buildup during the production of nasals.

The first subject, hereafter referred to as Subject 1, pronounced three repetitions of each utterance, for a total of 24 utterances. The second subject, hereafter referred to as Subject 2, pronounced eight repetitions of each utterance containing the vowel /ɑ/

and four repetitions of each utterance containing /i/, for a total of 96 utterances. The third subject, hereafter referred to as Subject 3, was the only subject to also record the voiced stops and the alveolar nasal. Subject 3 pronounced ten repetitions of each utterance, for a total of 260 utterances. The utterances were spoken in random order by each subject. A few repetitions had to be eliminated due to mispronounciation, slightly reducing the number of repetitions available for some utterances compared to the total number of repetitions discussed.

3.2 Recording Method

To transduce articulatory displacements, an electro-magnetic midsagittal articulometer (EMMA) was used. The application of EMMA to the investigation of stop-consonant production in this thesis is believed to be the first study of its kind, as no similar studies were located in the extensive literature search. A complete description of the development and operation of the EMMA system is found in Perkell et al., 1992. Briefly describing the device, three transmitter coils are mounted in an apparatus that is placed on the subject's head (Figure 3-1). The head is positioned within the apparatus such that the subject's midsagittal plane is aligned with the midline of the apparatus. Alternating magnetic fields are generated by the transmitters. The field strengths decrease approximately in proportion to the reciprocal of the distance cubed from the transmitter coils. Small, enclosed transducer coils (4 x 4 x 2.5 mm) are mounted with dental adhesive to articulatory structures of the subject near the midline of the apparatus. The transducers are positioned such that their axes are both as close to perpendicular to the midsagittal plane and as close to parallel to the transmitter axes as possible. As shown in Figure 3-2, the transducers can be mounted on the tongue blade, tongue body, lower incisors, lips and possibly the velum. Transducers are also mounted on the bridge of the nose and the upper central incisors for a maxillary frame of reference. This choice for frame of reference corrects for vertical and anteroposterior movements of the head with respect to the transmitter coil assembly.

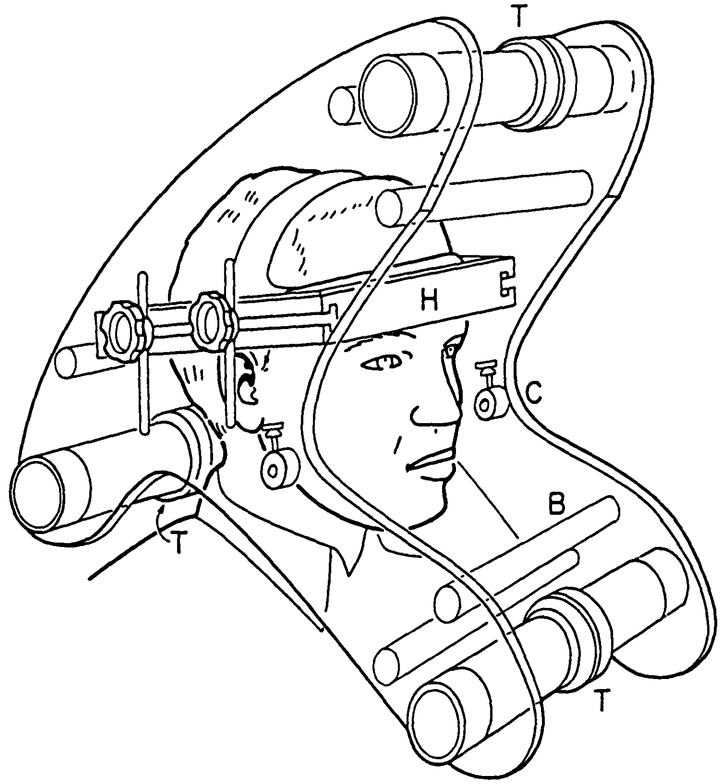

Figure 3-1: Electro-magnetic Midsagittal Articulometer (EMMA) system developed and used at the Massachusetts Institute of Technology. The transmitters (T) are held in place by side plates, connected by spacing bars (B). The headmount (H) fits precisely between the side plates. The center of the measurement area at the circular collar (C) is positioned over the cheek. Reprinted with permission from Perkell et al. (1992).

The transducer data utilized from the three experiments in this study (one experiment per subject) are from the lower lip (LL) and upper lip (UL) transducers placed on the vermillion borders of the lips, the mandible (M) transducer placed on the gingival papilla between the two lower central incisors, the tongue blade (TB) transducer, located on the midline of the tongue dorsum approximately 3 - 4 mm away from the tip of the tongue, and the TM transducer on the body of the tongue dorsum. The movement of the mandible is considered to be the same as that of the lower incisors, since the lower teeth are firmly rooted in sockets in the lower jaw. Consequently, the mandible transducer will be referred to interchangeably as the lower

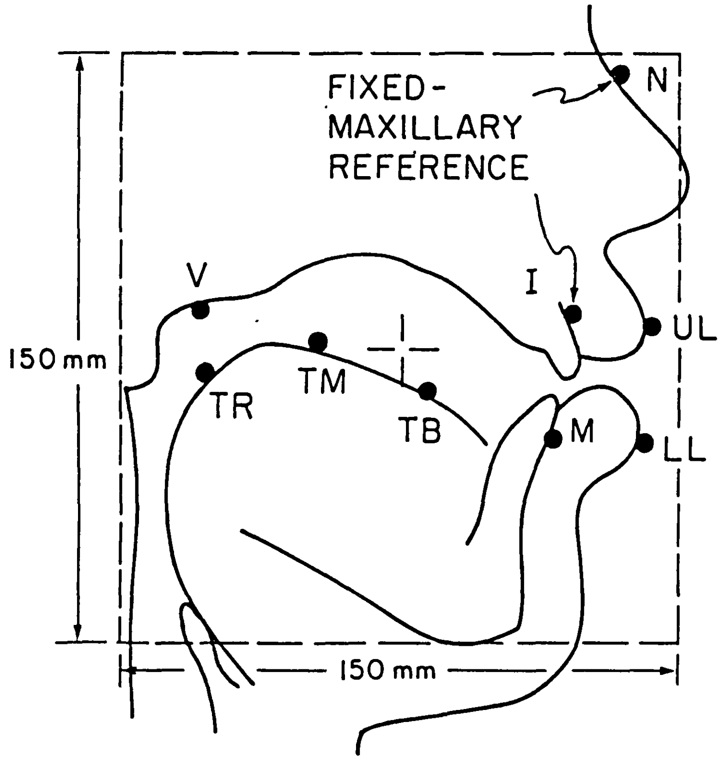

Figure 3-2: A schematic midsagittal view of a subject with nine possible measurement points (indicated by filled circles) corresponding to fixed points on the bridge of the nose (N) and the upper central incisors (I) for a maxillary frame of reference, and movable points on the mandible (M), lips (UL, LL), tongue (TB, TM, TR) and velum (V). Reprinted with permission from Perkell et al. (1992).

incisor (LI) transducer in this study. In addition, the tongue blade (TB) transducer will be interchangeably referred to as the tongue front (TF) transducer.

As the articulators move, measurement error can arise within this system from rotational misalignment of the transducers with respect to the transmitter axes. The system design, in combination with a signal processing algorithm, corrects for this misalignment; however, when the transducers are placed a significant distance (more than 3 mm) away from the midline or when the rotation becomes too great, the effectiveness of the apparatus and algorithm is greatly diminished (Perkell et al., 1992). The problem of rotational misalignment is greatest for the transducers placed on the tongue because its surface is capable of undergoing large changes between articulatory configurations. In order to guard against rotational misalignment in each of the three experiments, the transducers were initially placed properly and their positions were repeatedly verified throughout the experiments. This verification process resulted in the detection of a detached tongue blade (TB) transducer during the experiment for Subject 2, consequently the data recorded from that transducer during that experiment will not be analyzed.

During the experimental session, the subject was seated in a straight-backed chair with armrests and footrest. After placement of the transmitter coil assembly and transducers, the operation of the apparatus was checked via short trial recordings. The subject was instructed to speak at all times in a natural, fluent manner while avoiding unnecessary head movement. The utterances for the experiment were placed on a list directly in front of the subject.

The transducers are each connected by a fine twisted-wire pair to receiver electronics. [1] While the subject pronounces the utterances, the electronics convert the induced high frequency signals to slowly varying voltages corresponding to the distances from the transmitters. The voltages are digitized using an A/D converter, at a sampling rate of 312.5 Hz, and stored on disk. Signal processing software (Henke, 1989; Perkell et al., 1992) is then used to demultiplex the digitized signal into separate, synchronized articulatory signal streams and to convert the transducer voltages to x and y displacements in the midsagittal plane. The velocity and acceleration are also calculated for each transducer. The movement of the displacement coordinates is determined relative to a fixed reference point within the vocal tract, as established by the two fixed maxillary reference transducers. The measurement resolution is estimated to be better than 0.5 mm for the lip and lower incisor movements and better than 1 mm for the tongue movements (Perkell et al., 1992). The signal processing software includes a lowpass, smoothing filter to remove noise from the data. The filter has a cutoff frequency of approximately 12 - 16 Hz (the cutoff value varied slightly with each experiment). In Section 3.4, the most rapid velocities recorded by EMMA

[1] The connecting wires are believed to have a negligible effect on the movement of the articulators after a short period of accommodation by the subject.

are the lower lip and tongue blade velocities, in the range 20 - 30 cm/sec for an individual repetition. The range of the peak velocities is consistent with the peak velocity values reported in Chapter 1, Section 1.2.1, as recorded by other devices. Consequently, the cutoff frequency of the lowpass filter is believed to be high enough to preserve the most rapid movements of the articulators.

3.3 Displacement Analysis

First, the data processing procedure developed to average the displacement data is presented. An averaged displacement waveform for a given utterance and speaker is more representative of the speaker's articulatory movements than any single repetition of the given utterance. The averaging method takes into account slight variations in speaking rate between repetitions, via the application of linear time warping, so that much of the detailed signal structure of the movements is retained.

Second, the averaged displacement waveforms are analyzed. The coordination, timing, sequence and rates of movements for the voiceless labial and alveolar stops, in the context of various phonetic environments, will be discussed. Production of the voiceless stops will be compared to production of the voiced stops and nasals. Comparisons will be made to related findings cited in Chapter 1, Section 1.2.1, Articulation Experiments.

3.3.1 Data Processing Procedure

An example of the articulator displacement, time-aligned with the acoustic data, is given in Figure 3-3. For information on the recording of the acoustic data, refer to Chapter 4, Section 4.2.

At the completion of each EMMA experiment, the articulation and acoustic data were stored on a VAX computer and made accessible through Scanx, a software program written in the MITSYN language (Henke, 1989). The first step in the data processing performed by the author was to move the articulation data from the VAX to a UNIX computer for use with MATLAB, version 4.2c, a graphics and signal

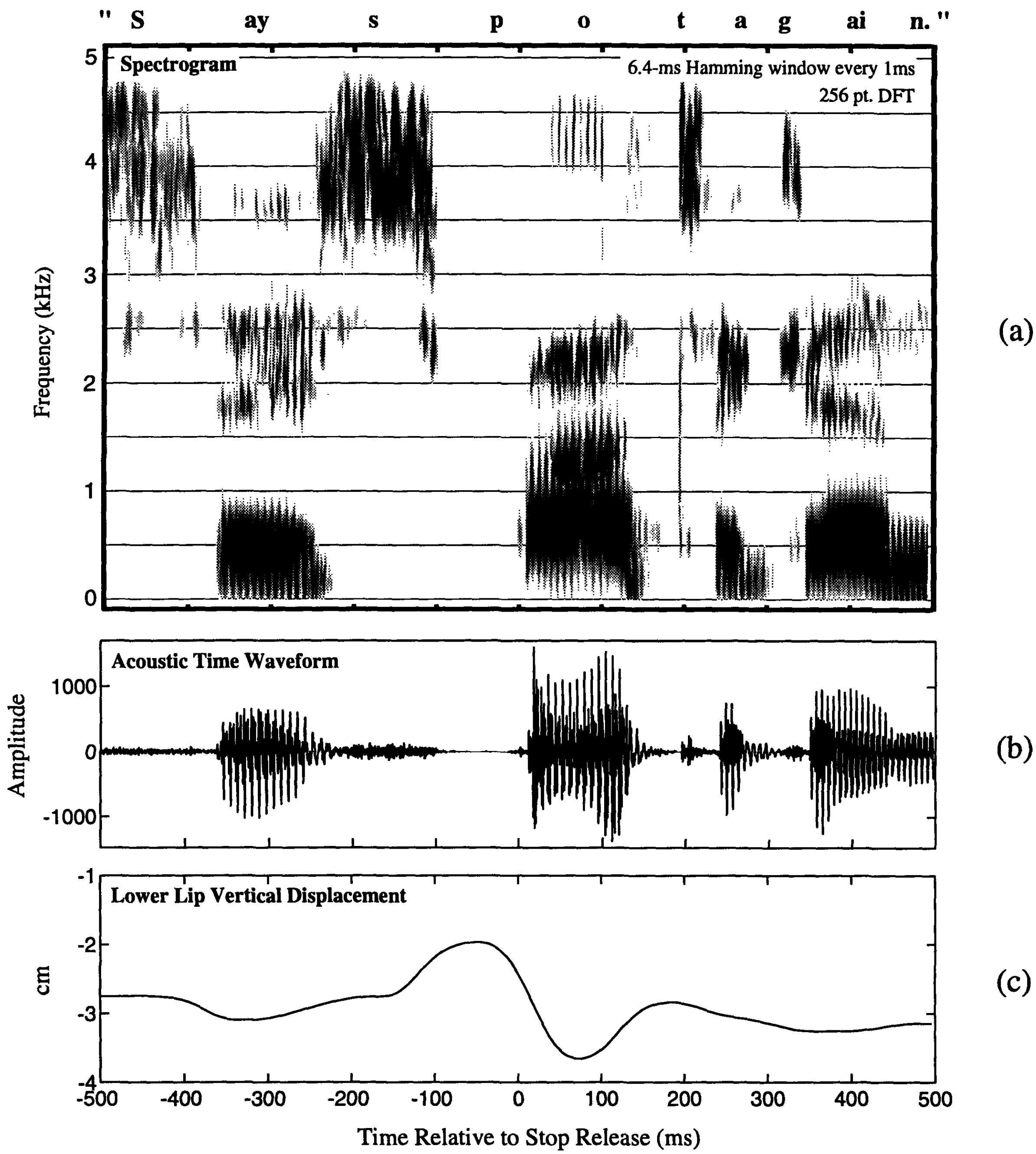

Figure 3-3: Example of articulation and acoustic data recorded during an EMMA experiment. One repetition of "Say spot again." spoken by Subject 2. (a) Spectrogram (produced using LSPECTO). (b) Acoustic time waveform (Amplitude is proportional to sound pressure recorded by microphone). (c) Lower lip vertical displacement, as measured by LL transducer relative to reference transducers.

processing software package by The MathWorks, Inc. MATLAB programming code was developed, as described below, to further process the articulation data. In order to analyze the acoustic data, the acoustic signals for each utterance were moved to a different VAX machine cluster, one on which KLSPEC and LSPECTO were available. KLSPEC and LSPECTO are software packages developed by D. H. Klatt, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA.

It was desirable to develop a data processing procedure which would average the articulation data across repetitions of a given utterance for a given speaker, while retaining as much of the detailed signal structure of the movements as possible. Averaged data would be more representative of a given speaker's articulator movements for a given utterance than any single repetition. However, a method for averaging had to be developed which took into account the slight variations in speaking rate between repetitions, such that the details of the articulator movements were not smoothed out during the averaging process. Briefly, the data processing procedure involves the identification of points in the articulation data when certain acoustic events occur in a given utterance, followed by a three-step averaging process of the articulation data, utilizing the identified articulation data points and involving the application of the signal processing technique of linear time warping. The specifics of the processing procedure are described in the remainder of this section.

The time-aligned acoustic waveform was utilized to determine points in time when certain acoustic events occur in the articulator movements. First, the event times were determined from the acoustic data. Then, the corresponding times and magnitudes were identified in the articulation data. The acoustic events include the end of the vowel /e/ in <u>say</u>, the end of the /s/ noise (if the fricative is present), the stop or nasal consonant release, and the onset of the following vowel /ɑ, i/. The total number of acoustic event points for a given utterance ranges from two to four points. A pictorial representation of the selection of the acoustic events in the articulation data appears in Figure 3-4. The end of the vowel /e/ in <u>say</u> was determined to be at the end of the last recognizable glottal pulse during production of the vowel. A recognizable glottal pulse is one which has the same shape, number of peaks and duration that

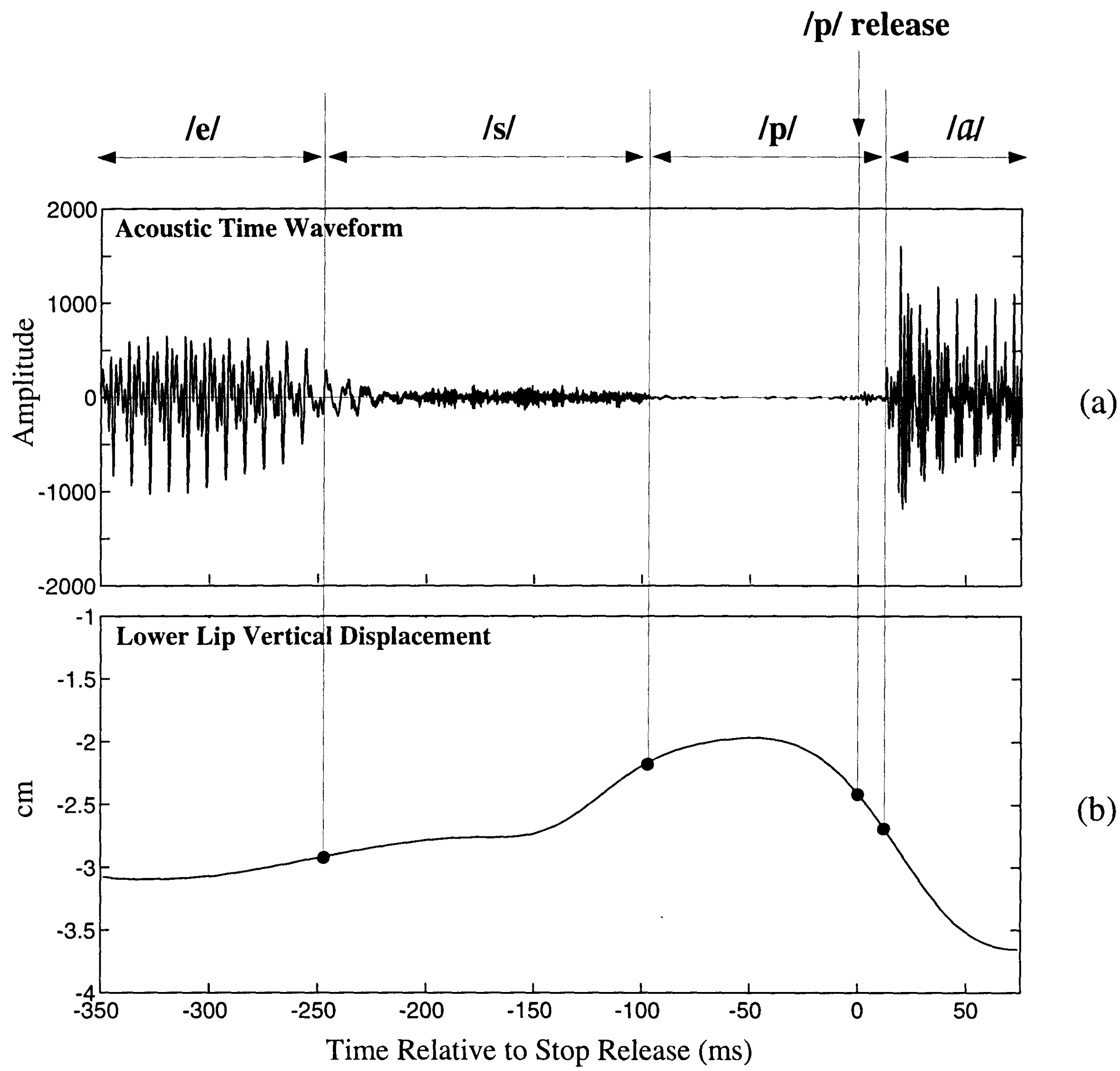

Figure 3-4: Determination of particular points in time in the articulation data when specific acoustic events occur. Excerpt from one repetition of "Say spot again." spoken by Subject 2. (a) Acoustic time waveform. (b) Lower lip vertical displacement. Acoustic event times are marked by vertical lines extending from their occurrences in the acoustic signal to end in filled circles at the corresponding points in time in the articulation data.

the majority of the glottal pulses have for that particular vowel. The end of the last recognizable glottal pulse is defined to be the positive zero crossing of the first glottal pulse which is no longer recognizable. The end of the fricative /s/ is fairly easy to determine since the frication noise ends abruptly. The particular point in time when the amplitude of the frication noise has died down to that of the background noise is defined to be the end of the /s/ production. The stop-consonant release is defined to be the point at which the amplitude abruptly increases from the level of background noise during consonantal closure to the level of noise produced during the transient burst at the release. The initiation of the transient burst upon release of the stop denotes the time when the constriction is just beginning to open. For /p, b/ the release is often difficult to detect. Due to the lack of a cavity in front of the constriction, the amplitude changes are very subtle, and information about the typical voice onset time (VOT) for the following vowel is useful in determining the general region in which to look for the /p, b/ release. The VOT is the time interval measured from the onset of burst release to the point where the time waveform first shows signs of periodicity following the release. If the /p, b/ release is unable to be detected, a point in the vicinity of the release is chosen, based on the expected duration of the VOT. Zue (1976) reported the mean VOT to be 13 msec for /b/, 15 msec for unaspirated /p/ and 58.5 msec for aspirated /p/. For /t, d, k, g/ detecting the release is ordinarily not a problem; however, multiple bursts may occur, especially for the velar stops. Multiple bursts involve the tongue moving downward to release the stop, then upward as if making the closure again, then downward for a second "release". It is possible to have two or more such "releases" in this way, generating multiple bursts. The cause of the multiple bursts can be attributed to the Bernoulli effect. The burst occurring first in time is chosen unless the constriction appears to completely close again (seen by the absence of any sound in the acoustic time waveform), then the burst immediately after the final such closure is chosen. If the constriction almost completely closes, such that the first noise burst has an amplitude only one-quarter or less than that of a nearby, later burst, then the later, larger burst is chosen. The nasal consonant release occurs simultaneously with the onset of the following vowel. The release is typically fairly easy to distinguish since there is a marked difference in appearance between the virtually uniform glottal pulses during the consonantal closure and the pulses of the following vowel. The particular point in time corresponding to the nasal release is determined in the same manner as the onset of the following vowel /ɑ, i/. The procedure to find the onset of the vowel /ɑ, i/ is to first locate the positive zero crossing of the first peak of the first vowel glottal pulse. Then the time before that first peak, when voicing initially appears to start (identified by the presence of periodic peaks, although not having the same period as the glottal pitch period, nor necessarily having the same shape) is defined to be the onset of the vowel. The time of the start of the vowel is not chosen necessarily to be a zero crossing, but rather to be the time when a significant increase in amplitude occurs. In the case of the vowel onset following a nasal consonant, the time is the point at which the glottal pulses change dramatically in appearance, as stated earlier, and does not necessarily correspond to an increase in amplitude, because the amplitude is already quite large during the closure interval. The acoustic event times were identified by the author on an utterance-by-utterance basis. The criteria discussed for determination of the acoustic event times were used as guidelines; the author's judgment was also utilized whenever necessary.

It was required that the averaging approach preserve the time and magnitude of the acoustic event points in the articulation data for a given utterance, as well as preserve the shape of the displacement waveform inbetween events. The first step was to average, across repetitions, the times and corresponding magnitudes of each acoustic event point in the articulation waveforms, obtaining a set of averaged acoustic event points. A demonstration for the simple case of averaging together two lower lip vertical displacement waveform repetitions appears in Figure 3-5.

The second step in the data processing procedure was to utilize linear time warping to retain as much detail of the displacement waveform shape as possible inbetween acoustic event points in the averaged waveform. For example, if N equally-spaced points were desired between two consecutive, averaged acoustic event points, then N equally-spaced points were located between each of the corresponding original

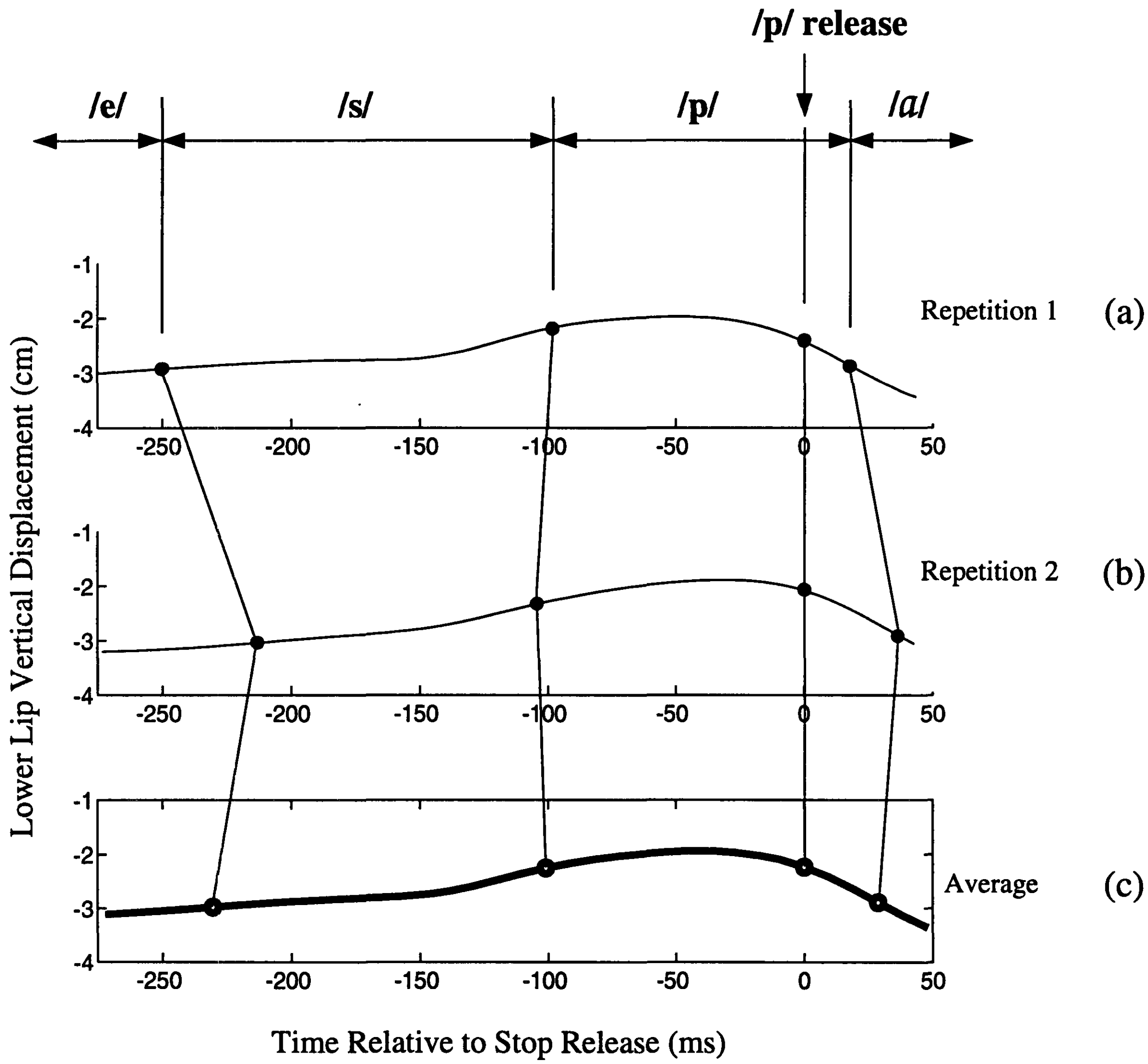

Figure 3-5: Demonstration of data processing averaging procedure, involving linear time warping. Excerpt from two repetitions of "Say spot again." spoken by Subject 2, and their computed average displacement waveform. (a) First repetition of lower lip vertical displacement. (b) Second repetition of lower lip vertical displacement. (c) Averaged lower lip vertical displacement waveform.

acoustic event points in each of the repetitions. These N points were then averaged across repetitions in the same manner as the acoustic event points, i.e. the times and magnitudes of the first point following each original acoustic event point were averaged together, respectively, and so on for the additional N - 1 points. If a point was not located at the selected sample location in a given repetition, then the nearest data point was chosen, its magnitude was retained, and its time was changed to the time of the sample location. Individual data points will move at most 1.6 msec with this approach. In order to avoid duplicating points, N should be selected to be greater than or equal to the number of data points which lie between the acoustic event points in the repetition with the shortest amount of time between those two consecutive, acoustic event points. This method results in the use of only actual data points, and does not perform interpolation between points. The approach yields an averaged displacement waveform, with N equally-spaced points between each two consecutive, averaged acoustic event points. The value of N can vary between different sets of consecutive averaged acoustic event points in a given utterance. The value depends upon the quantity of time between any two consecutive points. The more time, the larger the value of N. The nominal spacing for the N points is 10 msec between any two consecutive points. Due to the amount of time between any two consecutive points not being the same across repetitions, the technique described here is referred to as "time warping". The effects of time warping are able to be observed in Figure 3-5. The length of time between any two consecutive acoustic event points in the first repetition is slightly different from the length of time between the same two consecutive acoustic event points in the second repetition.

The third and final step in the data processing procedure was to extend the averaged displacement waveform for a length of time prior to the first averaged acoustic event point and after the last averaged acoustic event point. To include a few data points prior to the first averaged event point, the points lying approximately 10 msec before the first acoustic event point in each of the repetitions were extracted and averaged together. Next, the points approximately 20 msec prior to the first acoustic event point in each of the repetitions were averaged. The process continued as described, in 10-msec increments, for typically five to ten increments, providing several averaged data points prior to the first averaged acoustic event point in the averaged displacement waveform. In a similar fashion, averaged points were added to the end of the averaged displacement waveform, following the final averaged acoustic event point. As in the second data processing step, if an actual data point was not located at exactly a 10 msec-increment interval prior to the first acoustic event point or after the last acoustic event point, then the nearest actual data point was chosen. Consequently, the averaging technique for the third step also uses only actual data points, and performs no interpolation. The averaging technique described here does not "distort" or "warp" the time axis, thus it is not considered to be "time warping". This technique does not preserve the shape of the displacement waveform as well as time warping does, thus should not be used for more than a minimal length of time at either end of the averaged waveform, if close examination of the waveform shape is desired.

3.3.2 Results and Discussion

The graphs presented in this section all have a similar format. The x-axis is the time relative to the release of the stop or nasal consonant. A negative time corresponds to a time prior to the release. The y-axis is the average vertical displacement of the articulator(s) in the midsagittal plane, as recorded by the transducer(s). The displacement is measured in centimeters relative to a fixed vocal-tract reference point. The reference point is established by two fixed maxillary transducers, as described in Section 3.2. The acoustic event points, averaged as in Step 1 of Section 3.3.1, are labeled and represented by open circles. The displacement waveform between any two acoustic event points is the time-warped averaged displacement data. The time-warp averaging process is described in detail in Step 2 of Section 3.3.1. The waveform prior to the first acoustic event point and after the final acoustic event point in each displacement waveform is averaged using a simple averaging process described in Step 3 of Section 3.3.1. The graphs are grouped according to the observations made about them. Within each group, the x and y scales are the same, where possible, to facilitate comparison across graphs. The displacement waveforms of Subject 1 are typically not chosen for graphing, since the fewest number of repetitions per utterance (nominally three) were recorded for this subject.

Examination of the jaw movements in Figures 3-6, 3-7 and 3-8 reveals systematic differences between the production of voiceless stop consonants in word-initial position and voiceless stop consonants preceded by the fricative /s/. The production of /s/ places two constraints on the positioning of the articulators. The first constraint requires that the tongue blade be close enough to the hard palate to form a constriction, and the second constraint requires that the lower incisors be in a high enough position to create an obstacle to the airflow, resulting in the generation of turbulence noise downstream from the constriction. These constraints, particularly the latter, result in a high, fairly fixed position for the lower jaw throughout the production of /s/. Since the jaw is required to remain in a high position until the end of /s/, yet the position of the jaw for the following vowel /ɑ/ is also constrained (to be in a low position) the rate of jaw movement downward at the time of the release is rapid. For production of /p/ and /k/, in Figures 3-6 and 3-8, respectively, the maximum rate of downward jaw movement near the time of release is greater for the utterances containing /s/. The rate is 5.8 cm/sec for <u>spot</u> vs. 4.2 cm/sec for <u>pot</u>, and 5.2 cm/sec for <u>skot</u> vs. 2.9 cm/sec for <u>kot</u>. This trend also holds for the other two subjects. The constraints involved in the production of /s/ have influenced the production of the following stop consonant /p/ or /k/, resulting in a faster rate of downward jaw movement at the time of the stop release. This influence is referred to as "carry-over coarticulation". A similar observation can be made for production of the initial /t/ in <u>stot</u>, Figure 3-7, in which the maximum rate of downward jaw movement near the time of release is 8.0 cm/sec for <u>stot</u> vs. 6.5 cm/sec for <u>tot</u>. For the production of both /s/ and /t/, the jaw is close to the same high position, as seen in Figure 3-7. The high jaw position can be attributed both to the tongue blade forming the constriction in each case and to the need to generate turbulence noise behind the incisors during the production of each of the two sounds. Comparison with the other two subjects shows Subject 3 also has a faster rate for /t/ preceded by /s/, however Subject 1

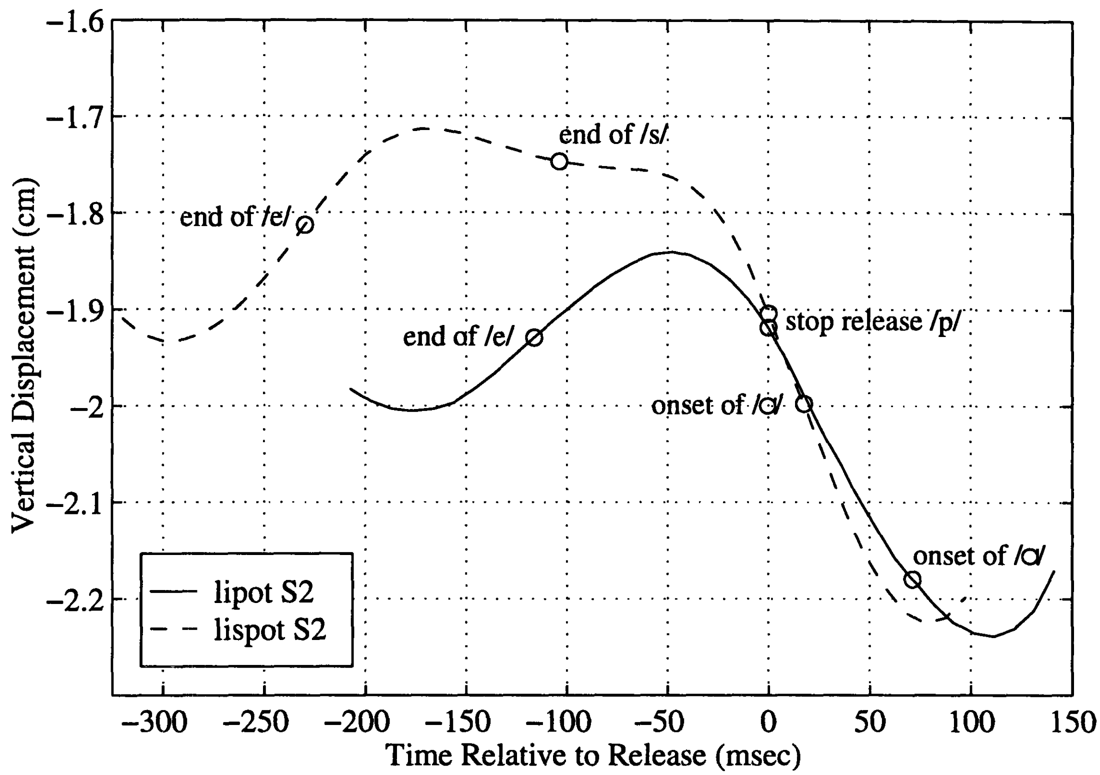

Figure 3-6: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterances pot and spot, spoken by Subject 2. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

shows very similar rates, with a slightly faster rate for the word-initial /t/, which is not unexpected, since the jaw position during the constriction is so similar in each case.

Comparing across subjects, Subject 3 consistently has the greatest jaw height difference prior to the release between the utterances containing the voiceless stops preceded by the fricative /s/ and the utterances with the voiceless stop consonants in word-initial position. For example, the utterances containing /s/ have a jaw position ranging from about 2 to 5 mm higher at the end of /s/ than the utterances not containing /s/ at a similar time prior to the stop release.

In addition to carry-over coarticulation, in which the context prior to the stop consonant influences the production of the stop, "anticipatory coarticulation" also

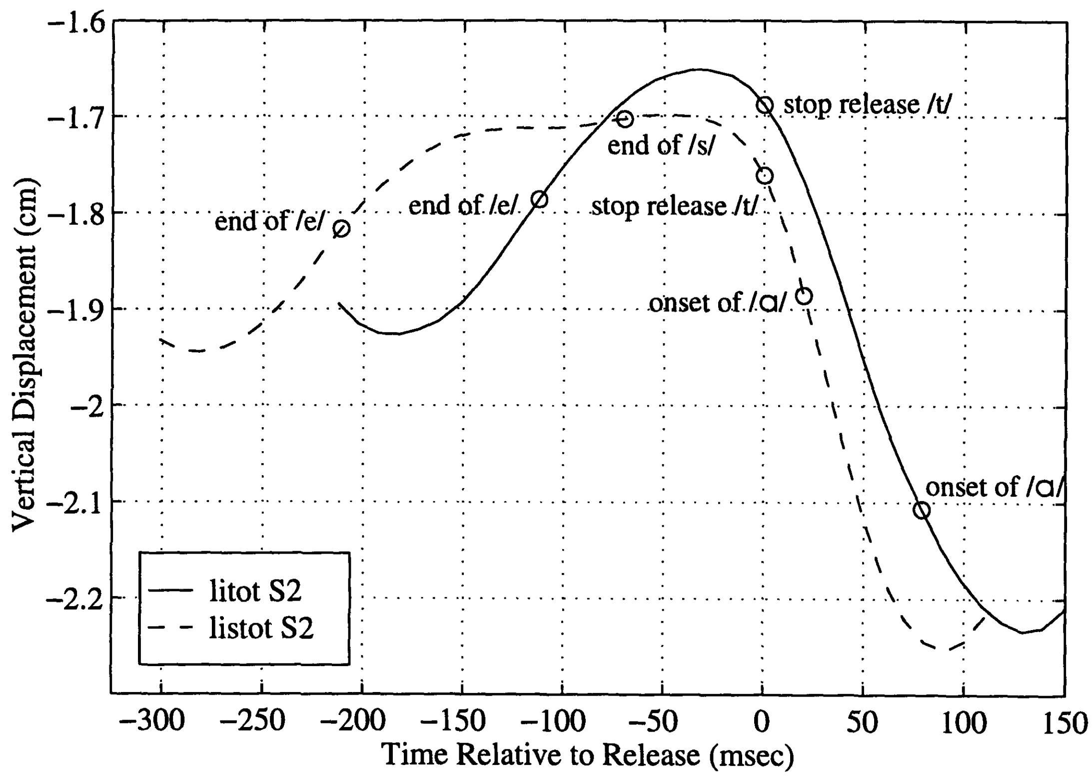

Figure 3-7: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the first /t/ stop release in the utterances tot and stot, spoken by Subject 2. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

influences stop production. Anticipatory coarticulation is the influence of the context following the stop on the production of the stop. The jaw position during production of /e/ preceding /p/ in pot and peet, Figure 3-9, differs by less than 1 mm. During the closure interval of /p/, however, anticipation of the following vowel is able to be observed. The vowel /ɑ/ requires a low, back position for the tongue body. The jaw position must be low to facilitate this tongue position. In anticipation of the low jaw position required to produce the following vowel, the jaw begins moving down during the /p/ closure interval. In contrast, the vowel /i/ requires a high, front tongue position, so the jaw needs to move only slightly downward in anticipation of this position. There is a difference in jaw height between the two utterances of almost 2 mm by the time the stop release time is reached. By the time the vowel steady-

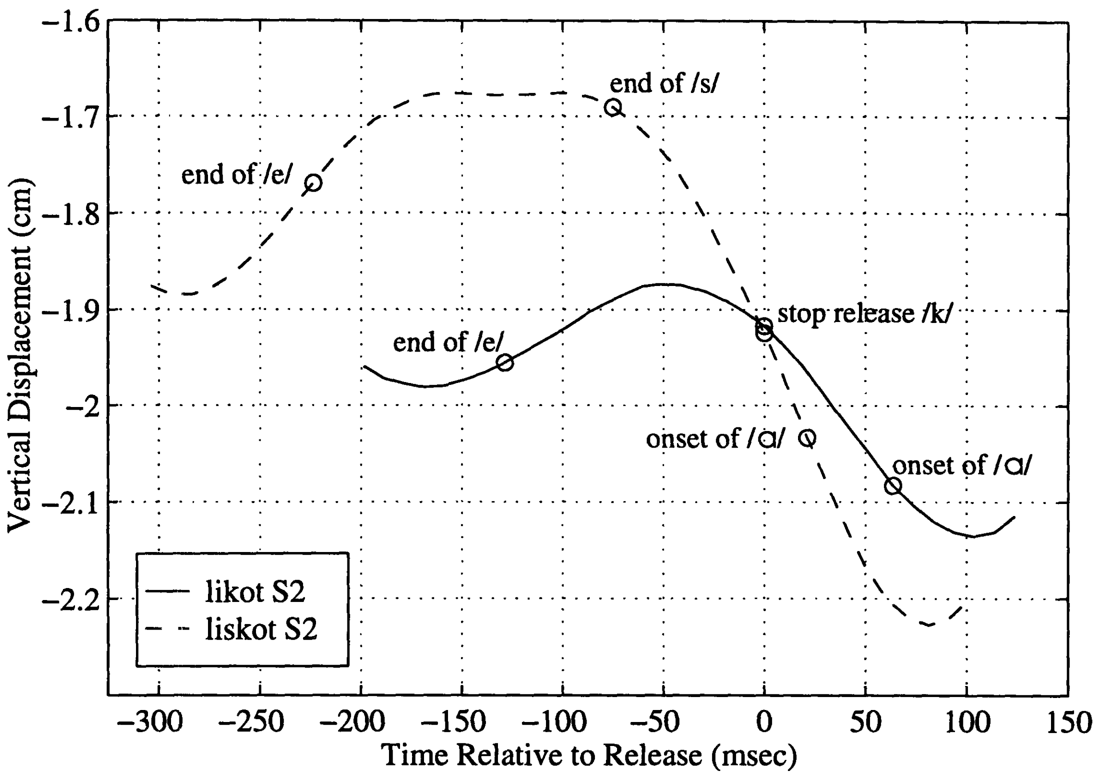

Figure 3-8: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the /k/ stop release in the utterances <u>kot</u> and <u>skot</u>, spoken by Subject 2. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

state is reached, there is approximately a 4 mm difference in jaw height. Similar observations were made for the other two subjects.

In order to reach the position required for the following vowel in a timely fashion, such that the amount of time between the /p/ release and the steady-state portion of the following vowel is similar for each utterance, the rate of downward movement of the jaw must be faster for the utterance containing /ɑ/. Indeed, the maximum rate of downward jaw movement near the time of the stop release is 6.3 cm/sec for <u>pot</u> vs. 1.2 cm/sec for <u>peet</u> (for Subject 3). Similar observations were made for the other two subjects.

Combining the observations that the jaw both moves downward more rapidly as well as moves downward farther in anticipation of the vowel /ɑ/ leads to the conclusion

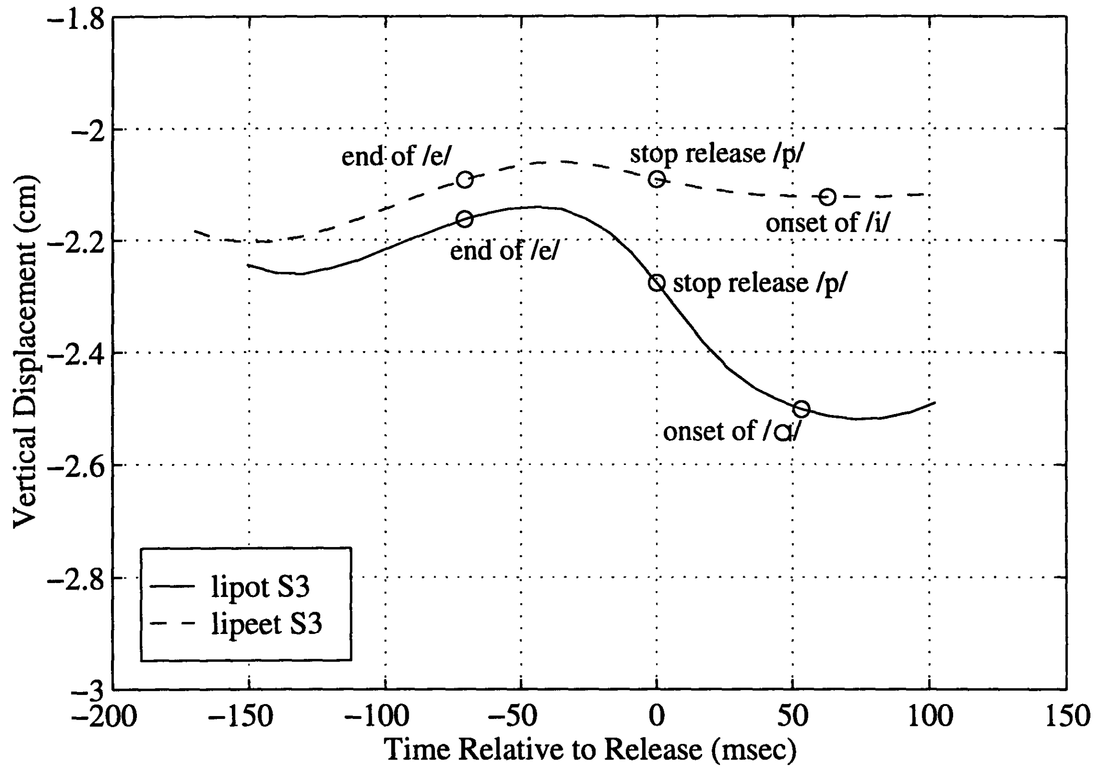

Figure 3-9: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterances pot and peet, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

that there is a correlation between increasing distance articulators are required to move and increasing rates of movement. This conclusion agrees with the observation, made by Kuehn and Moll (1976) and others, that a faster velocity corresponds to a greater displacement. For Subjects 1 and 2, a further observation can be made that, even though the rate of downward jaw movement has increased during the closure interval in anticipation of the following vowel /ɑ/, the rate is not fast enough to result in the steady-state portion of /ɑ/ occurring as quickly after the stop release as the steady-state portion of /i/. This observation may be evidence that production of the vowel /ɑ/ may not be anticipated as fully during production of /p/ as the vowel /i/.

The lower lip vertical movements for pot and peet, Figure 3-10, show only minimal difference in jaw height until after the stop release. Anticipatory coarticulation is not observed in the lower lip movement prior to the stop release. The lower lip is a primary articulator, responsible for forming the labial constriction, and is consequently not greatly influenced by production requirements of the following vowel. In contrast, the lower jaw is a secondary articulator, required only to assist in the formation of the constriction, leaving it more free to move into a position which partially anticipates the production of the upcoming vowel. The lower lip has a much faster maximum velocity following the stop release, 16.5 cm/sec vs. 6.3 cm/sec for the lower jaw in pot, and 11.6 cm/sec for the lower lip vs. 1.2 cm/sec for the lower jaw in peet for Subject 3. In addition, the amount of downward movement for the lower lip is greater than that of the lower jaw. For the lower lip, the vertical displacement between the position at the stop release and the position during the steady-state portion of the following vowel /ɑ/, is almost 8 mm. For the lower jaw, the displacement between the same two jaw positions is a little more than 2 mm. These trends are further indicators that the lower lip is a primary articulator, since the articulator primarily responsible for releasing the stop-consonant constriction should move the most and have the fastest velocity following the release. The trends regarding lower lip movement hold across all three subjects.

One aspect of the production of stop consonants is the buildup of pressure during the closure interval. Upon release of the stop, the intraoral pressure rapidly diminishes. In contrast, there is no similar intraoral pressure buildup during the closure interval for nasals because the velum is lowered, allowing the nasal cavity to act as a "pressure-release valve". The lower lip movements for utterances containing /p/ and /m/ were studied to determine if the effects of the pressure buildup and release were able to be recorded by the transducers. Figure 3-11 compares lower lip movements during production of the utterances pot and mot, for Subject 2. The lower lip trajectories are very similar throughout production of /p/ and /m/. If the transducer was able to record the release of the intraoral pressure, it might be expected that the maximum rate of lower lip downward velocity near the time of the stop release would be faster than for the nasal release. The pressure release would, in effect, "blow the lips apart" more rapidly than they would otherwise separate following the release.

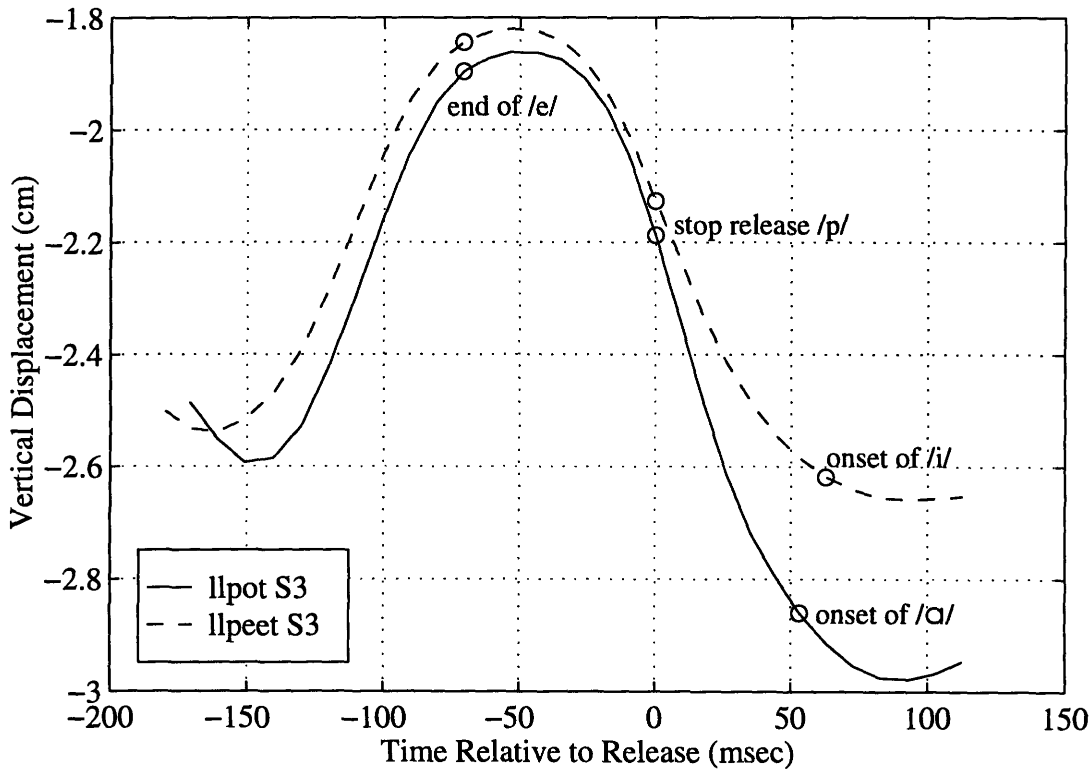

Figure 3-10: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterances <u>pot</u> and <u>peet</u>, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

However, only one of the three subjects recorded a faster maximum lower lip velocity for the stop consonant following release. The conclusion is that the lower lip transducer, placed at the vermillion border of the lower lip, is not placed in a position to capture all of the movement of the lower lip at the time of the release. In particular, the movement of the inner edge of the lower lip, which would be most likely to show the effects of the release of pressure, is not recorded by the transducer.

An interesting observation can be made across subjects. The lower lip trajectories for <u>pot</u> and <u>mot</u> consistently follow a pattern. Initially, the lower lip for /m/ is in a higher position (during the end of the production of the vowel /e/ and for the initial portion of the closure interval), then the trajectories cross, such that the lower lip for /p/ is in a higher position (during the remainder of the closure interval and

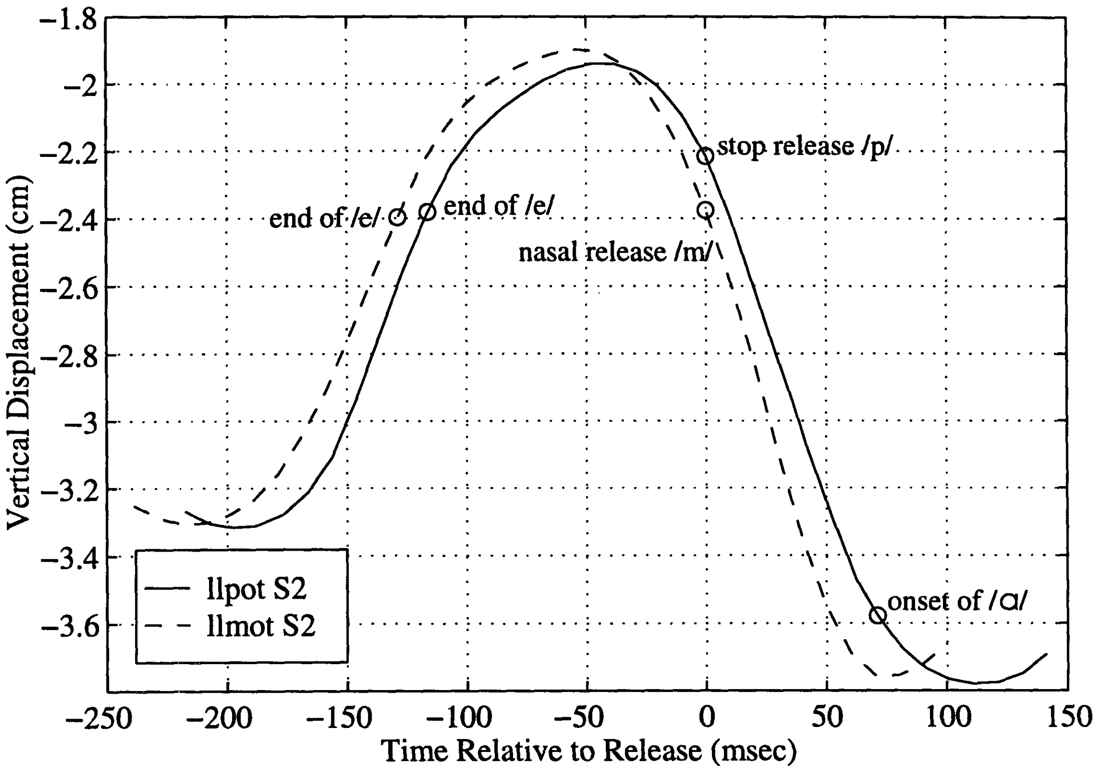

Figure 3-11: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop and /m/ nasal releases in the utterances pot and mot, respectively, spoken by Subject 2. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled. The nasal /m/ release and the onset of /ɑ/ occur simultaneously in mot.

throughout the time period following release). It is not thought that this pattern can be linked specifically to the presence or absence of pressure buildup, however, since the lower lip trajectories for peet and meet (Figure 3-12, for Subject 2) do not exhibit the same pattern for any of the subjects. (See discussion related to Figure 3-15.) The observation made earlier that the effects of pressure buildup cannot be seen in the transducer recordings holds for peet and meet as well, as demonstrated in Figure 3-12 where the trajectories are virtually identical throughout almost all of the closure as well as during the time period following release for the stop and nasal production.

The effects of presence or absence of pressure buildup can also be investigated for the alveolar stop and nasal consonants. Figure 3-13 shows the tongue blade

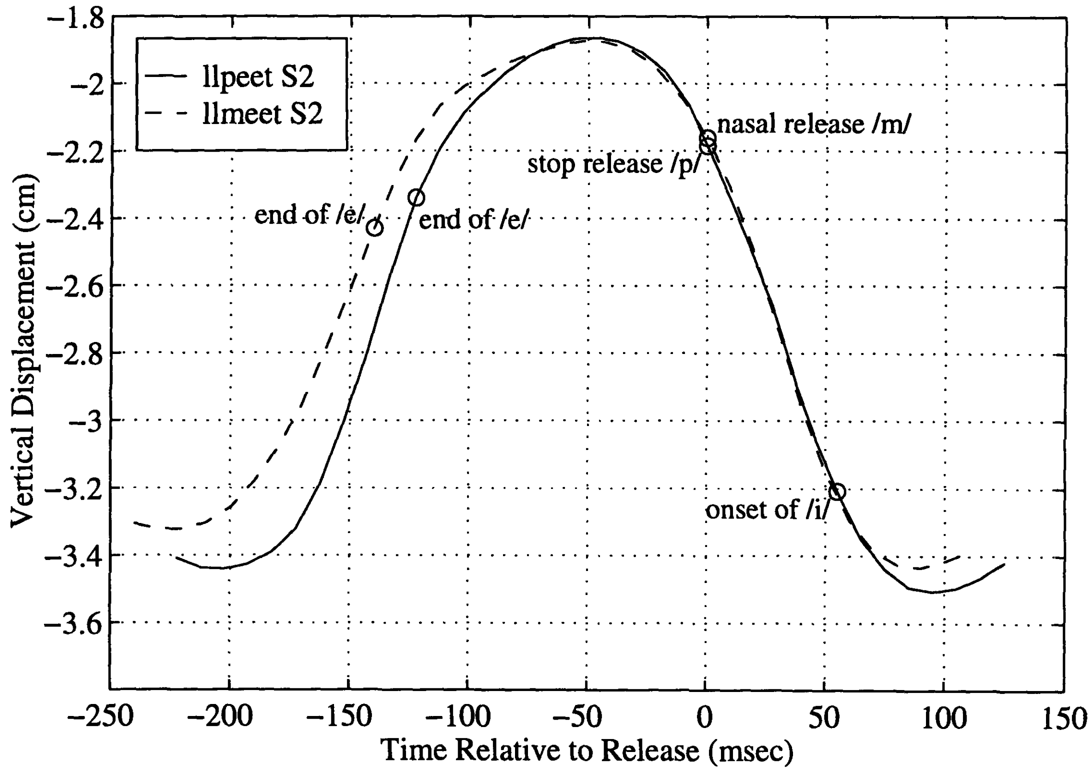

Figure 3-12: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop and /m/ nasal releases in the utterances peet and meet, respectively, spoken by Subject 2. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled. The nasal /m/ release and the onset of /i/ occur simultaneously in meet.

trajectories for the utterances tot, not and dot. The trajectories for the stops /t/ and /d/, in which pressure buildup occurs during the closure interval, are remarkably similar to one another. The trajectory for the nasal consonant /n/, for which there is no pressure buildup during closure, is noticeably different. During both the closure interval and the time period following release, the jaw does not reach as high a position for the utterance containing /n/. One possible explanation is that the jaw is deliberately held in a higher position during the closure interval for stop consonants in order to maintain closure against increased pressure behind the constriction, thereby preventing escape of air prior to the planned release time. This observation and explanation should be regarded as tentative, since data for the consonants /d/ and

/n/ was only available for one subject, although the average vertical jaw displacements were averaged across 9 - 10 repetitions per utterance.

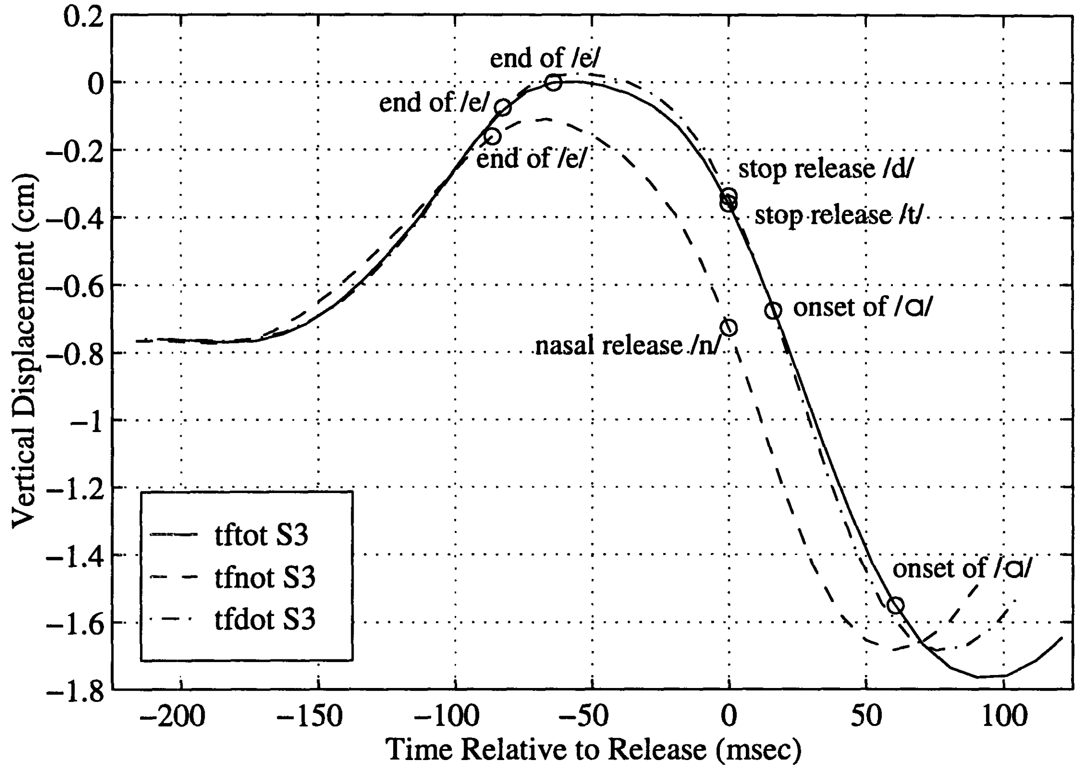

Figure 3-13: Average vertical displacement of the tongue blade, from a fixed reference point in the vocal tract, as a function of time relative to the initial alveolar stop or nasal release in the utterances tot, not and dot, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

A comparison of production of voiceless and voiced stops can be made from Figures 3-13, 3-14 and 3-15 for one subject. Virtually no differences are seen in the production of the alveolar stops, as noted earlier (Figure 3-13). Likewise, the velar stops are produced with virtually identical jaw heights (Figure 3-14). The labial voiceless and voiced stops, Figure 3-15, demonstrate the same pattern between their trajectories as was observed for the labial stop /p/ and nasal /m/ in the utterances pot and mot. Initially, the lower lip for /b/ is in a higher position (during the end of the production of the vowel /e/ and for the initial portion of the closure interval), then the trajectories cross, such that the lower lip for /p/ is in a higher position (during the remainder of the closure interval and throughout the time period following release). Although the trajectories are within measurement resolution (0.5 mm) of each other for the time interval when the lower lip is in a higher position for /p/, the trajectories of /p/ and /m/ were often within measurement resolution of one other as well. The observation that a similar trajectory pattern exists for the labial voiceless and voiced stops preceding /ɑ/ as existed for the labial voiceless and nasal consonants preceding /ɑ/ is further evidence that the pattern is probably not linked to the presence or absence of pressure buildup.

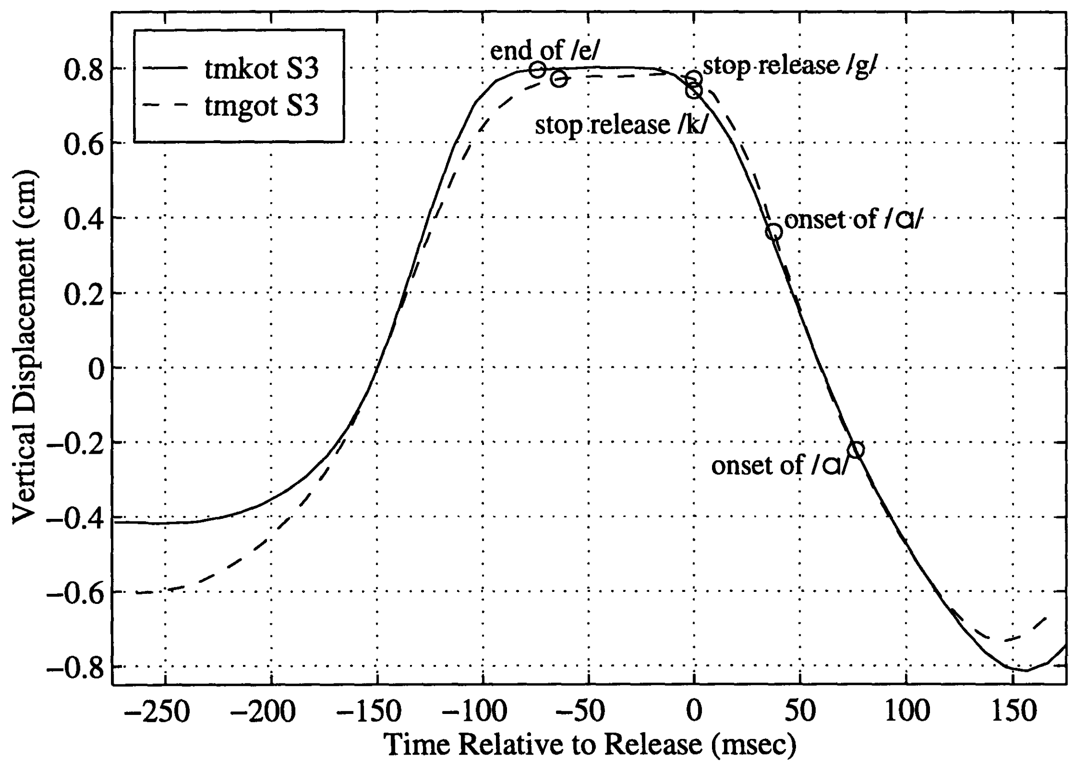

Figure 3-14: Average vertical displacement of the tongue body, from a fixed reference point in the vocal tract, as a function of time relative to the velar stop release in the utterances kot and got, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

Speakers often differ in their strategies for producing sounds. To observe variation across speakers, for the production of /p/ in a contextually-constrained environment, the average vertical jaw displacement during the production of spot for all three

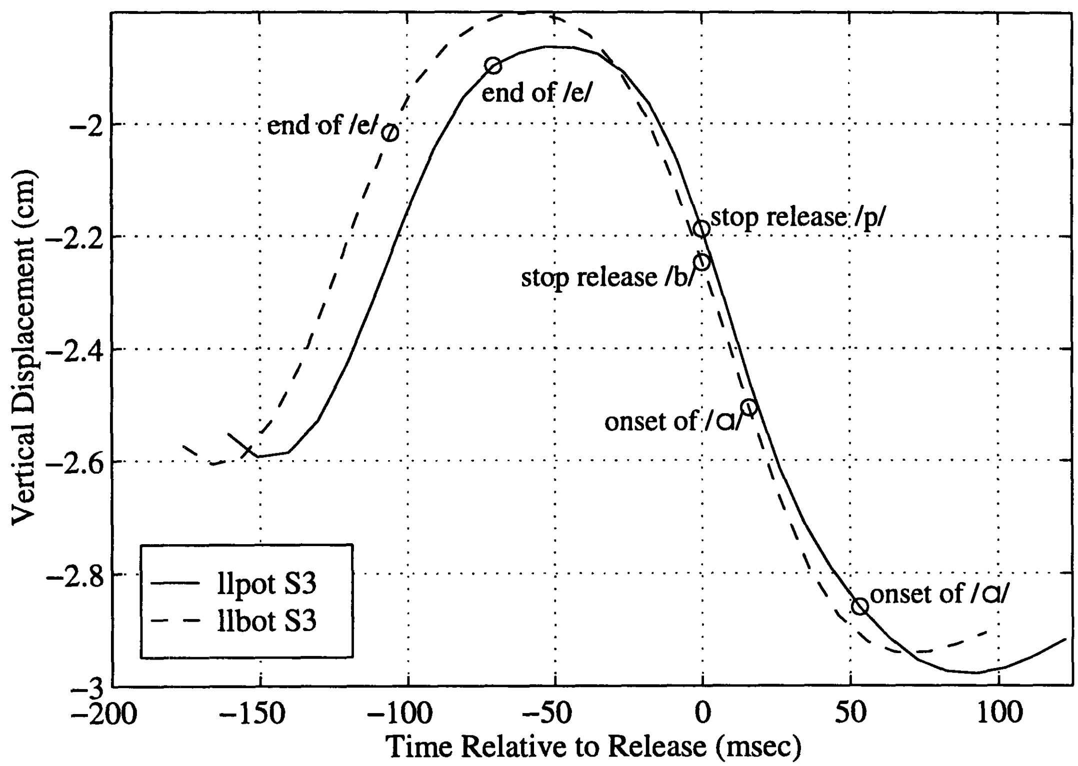

Figure 3-15: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the labial stop release in the utterances pot and bot, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

speakers is shown in Figure 3-16. As discussed in Section 3.1, the jaw position is fairly invariant across repetitions, contexts and speakers during the production of /s/, (Engstrand, 1980; Ichikawa et al., 1993). This knowledge leads to the observation from Figure 3-16 that the lower jaw transducer is placed on the lower incisors at close to the same position across speakers (within about 1 mm, measured at the end of /s/). Similar placement of lower jaw transducers across speakers allows for comparison of absolute jaw heights, in addition to comparison of rate and relative displacements.

Comparing maximum rates of downward jaw movement occurring shortly after the /p/ release in spot, Subject 1 had a rate of 8.0 cm/sec, Subject 2 had a rate of 5.8 cm/sec and Subject 3 had a rate of 7.7 cm/sec. The relative displacement from the end of /s/ to the steady-state portion of the vowel (when the trajectory reached

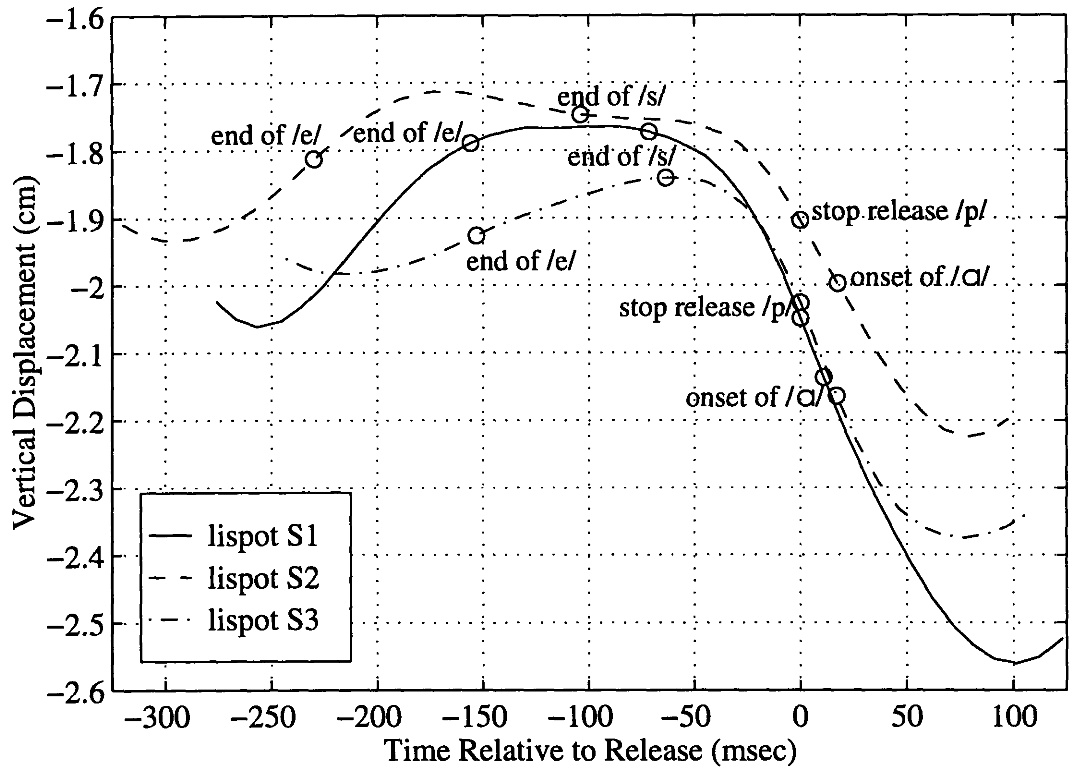

Figure 3-16: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterance spot, spoken by all three subjects. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

a plateau following the onset of /ɑ/) is about 8 mm for Subject 1, not quite 5 mm for Subject 2, and about 5 mm for Subject 3. The findings of Subject 1 and 2 can be explained by the observation that the amount of displacement is related to the rate of movement. Subject 1 has greater displacement and a faster rate than Subject 2, with a smaller displacement and a slower rate. Subject 3 has combined a faster rate with a smaller displacement, differing in strategy from the other two speakers for the production of /p/ in this context.

The lower jaw average vertical displacement during the production of pot for all three speakers is shown in Figure 3-17. First, comparing absolute height differences during the production of /p/, Subject 2 holds his jaw in the highest position, followed by Subject 1. Subject 3 holds his jaw in the lowest position of the three speakers. Considering the transducer position to be similar across speakers during /s/ production, thus aligning the transducer locations by removing the height difference seen at the end of /s/ in spot, there is approximately a 1 mm height difference between any two speakers' absolute jaw positions during /p/ production. The maximum rates of downward jaw movement occurring shortly after the /p/ release in pot are 6.8 cm/sec for Subject 1, 4.2 cm/sec for Subject 2 and 6.3 cm/sec for Subject 3. The relative jaw displacement, from the maximum height (reached during the closure interval of /p/) to the steady-state portion of the vowel (when the trajectory reached a plateau following the onset of /ɑ/), is a little more than 6 mm for Subject 1, about 4 mm for Subject 2 and a little less than 4 mm for Subject 3. Subject 2 once again has the slowest rate of the three speakers, whereas the rates for the other two subjects are more comparable. Subjects 1 and 2 again follow the displacement-rate correlation whereas Subject 3 again shows a fast rate associated with a small amount of displacement. The speakers each appear to have their own unique and consistent strategy for the jaw movement in production of the labial stop, even when the stop is placed in different contexts.

Across-speaker variations can also be examined for a transducer located on a primary articulator, such as the lower lip. The lower lip displacement was compared across speakers for the production of /p/ in pot, Figure 3-18. Comparing maximum rates of downward lower lip movement occurring shortly after the stop release, Subject 1 had a rate of 17.6 cm/sec, Subject 2 had a rate of 22.6 cm/sec and Subject 3 had a rate of 16.5 cm/sec. The relative displacement from the maximum height (reached during the closure interval of /p/) to the steady-state portion of the vowel (when the trajectory reached a plateau following the onset of /ɑ/) is about 15 mm for Subject 1, a little more than 18 mm for Subject 2, and about 11 mm for Subject 3. All three speakers' strategies conform with the displacement-rate correlation. Subject 2, with the largest displacement and fastest rate of movement, also takes the longest amount of time following the release to reach steady state for the vowel /ɑ/, possibly indicating a lack of complete compensation for the large displacement.

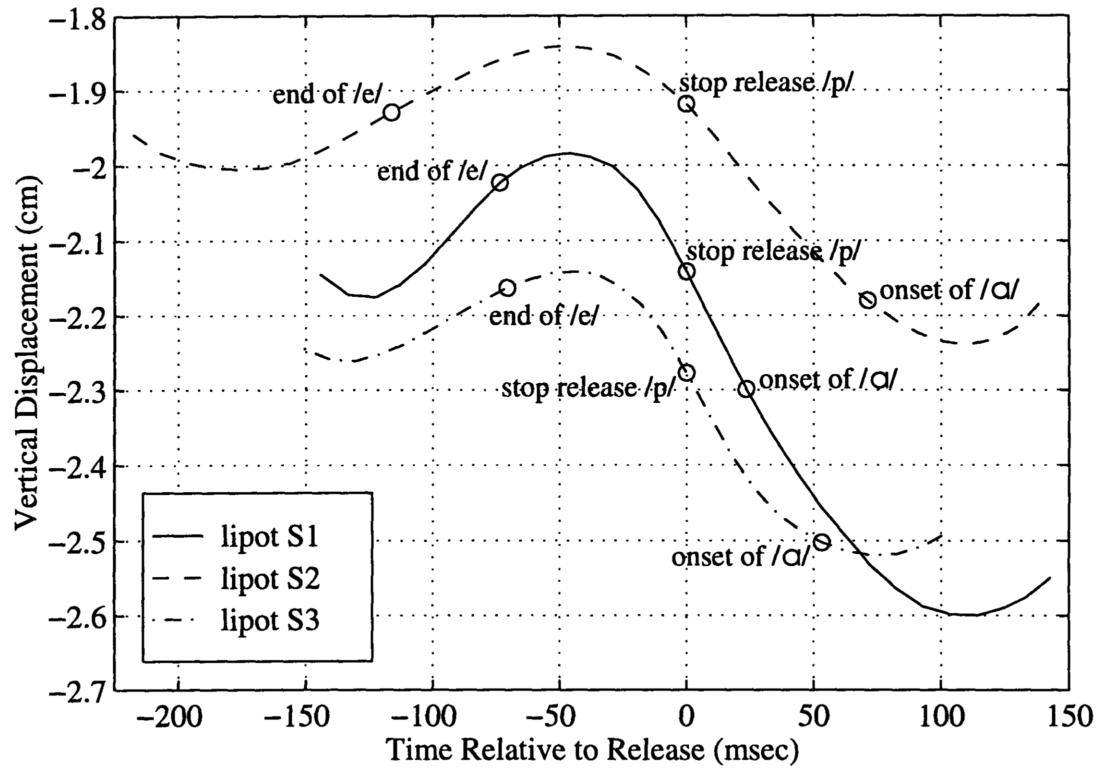

Figure 3-17: Average vertical displacement of the lower jaw, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterance pot, spoken by all three subjects. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

Further examination of Figures 3-17 and 3-18 reveals an apparent "cyclic" movement of the articulators from one configuration to another and back again. The cyclic movement is observed regardless of speaker or articulator. For example, the lower lip of Subject 1 in Figure 3-18 has a trajectory resembling a sinusoidal curve. The lower lip traverses through one complete "cycle" by starting approximately 130 msec prior to the release and ending about 110 msec or so following the release. The total duration of the lower lip cyclic movement is about 240 msec. Similar measurements of the cycles for Subjects 2 and 3 are approximately 310 and 230 msec, respectively. According to Stevens (Acoustic Phonetics, in preparation), the alternating closing and opening movements of the lips require a minimum time of 150 - 200 msec for

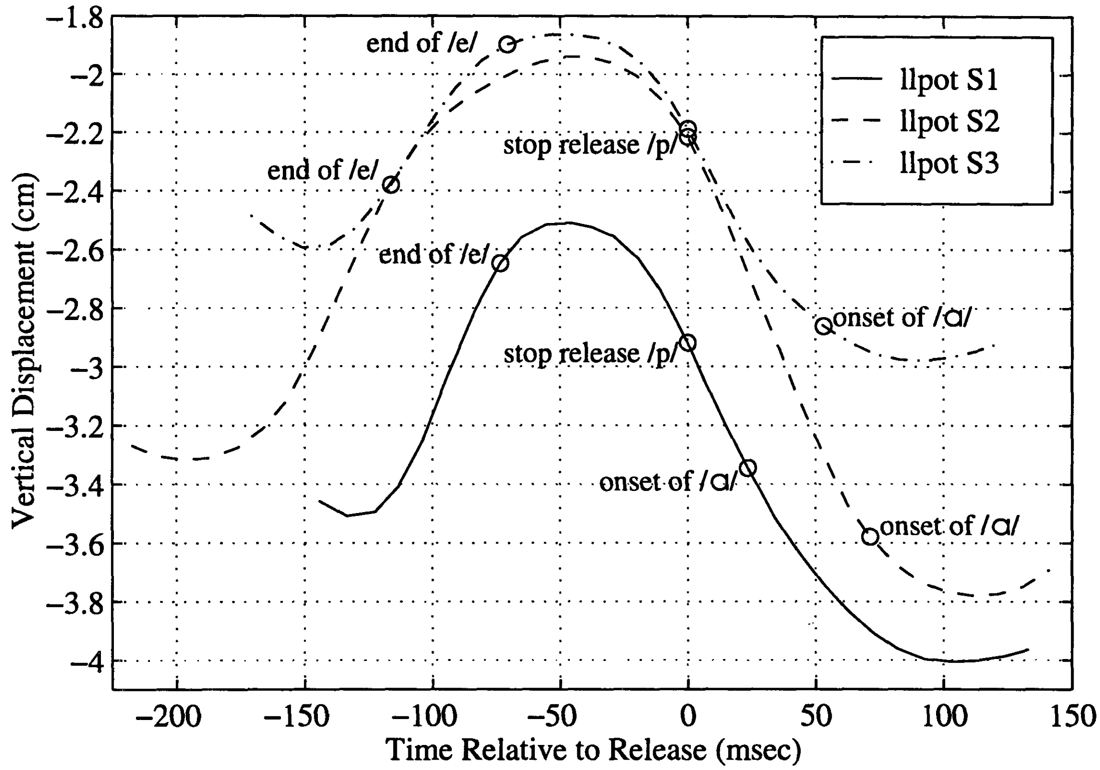

Figure 3-18: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ stop release in the utterance pot, spoken by all three subjects. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

a complete cycle, corresponding to a bandwidth of 5 - 6.7 Hz. For the lower jaw in Figure 3-17, the cycle is about 240 msec for Subject 1, 285 msec for Subject 2 and 210 msec for Subject 3. Nelson, Perkell and Westbury (1984) found a maximum frequency of 7 Hz for alternating movements of the jaw in repeated-syllable speech. The bandwidth of the cyclic movements for the lower lip and jaw in the present study is in the range 3.2 - 4.8 Hz, somewhat shorter than the maximum bandwidths. It is observed that Subject 2 has a slower cycle for the movement of each articulator than the other two speakers.

Additional observations regarding production can be made with the aid of Figures 3-19 and 3-20. The tongue blade is the primary articulator during the production of /t/ in tot, whereas there is no tongue involvement in the production of /p/ in pot. The movement of the tongue blade can be seen to anticipate the production of /t/

as early as about 170 msec prior to the release in Figure 3-19, evidence of planning well in advance to assure that the tongue blade will be in a position to execute the stop release at the desired time. The articulator movements are relatively slow, and planned for well in advance of the release, in contrast to the rapid acoustic changes executed at the time of the release.

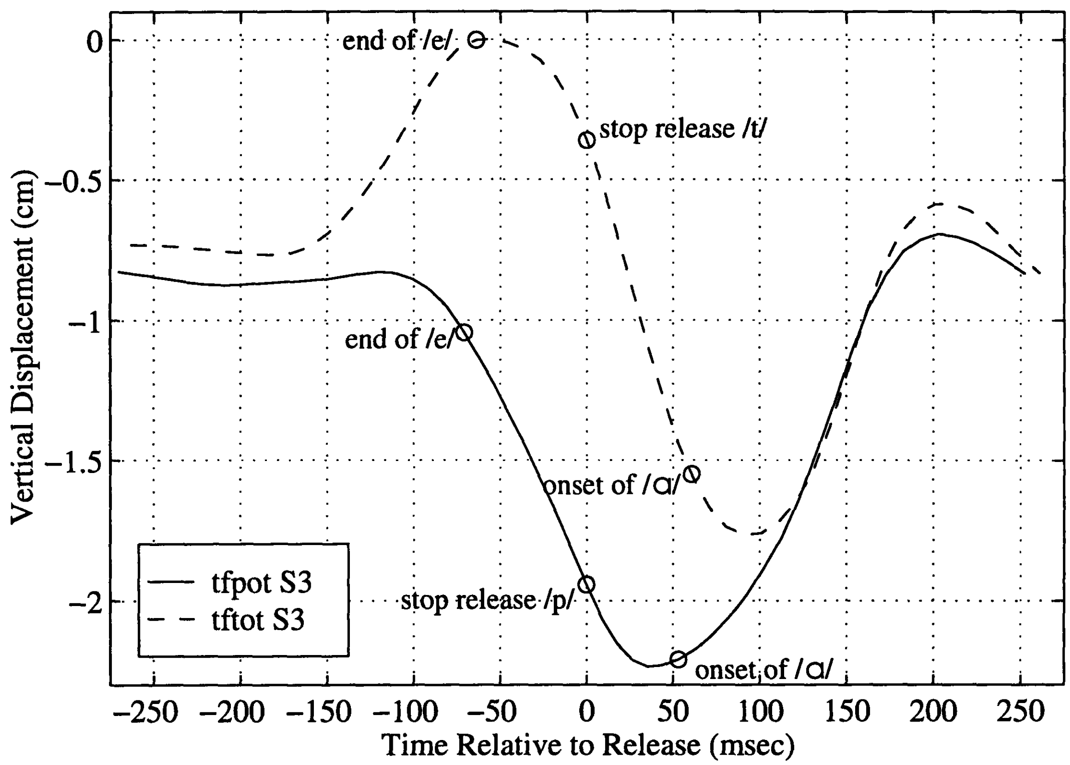

Figure 3-19: Average vertical displacement of the tongue blade, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ and initial /t/ stop releases in the utterances pot and tot, respectively, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

The time period during which the tongue blade is involved in the production of /t/ (when the tongue blade trajectories differ in movement in Figure 3-19) includes all three phases of the stop production. The time period can be seen in Figure 3-19 to extend from about 170 msec prior to the stop release to about 120 msec after the release, for a total duration of 290 msec. The duration is the time period of a cyclic movement (as discussed in the previous paragraph).

Similar observations can be made for the lower lip involvement in the production of /p/ in pot compared to the production of /k/ in kot, in which there is no constraint on the lower lip movement. Examination of Figure 3-20 reveals lower lip upward movement, toward /p/ closure, begins as early as 80 msec prior to the end of /e/, as much as 150 msec prior to /p/ release. The total time interval of lower lip involvement during /p/ production is 230 msec, as observed earlier. Similar observations were made for the tongue blade and lower lip movements in the other two subjects.

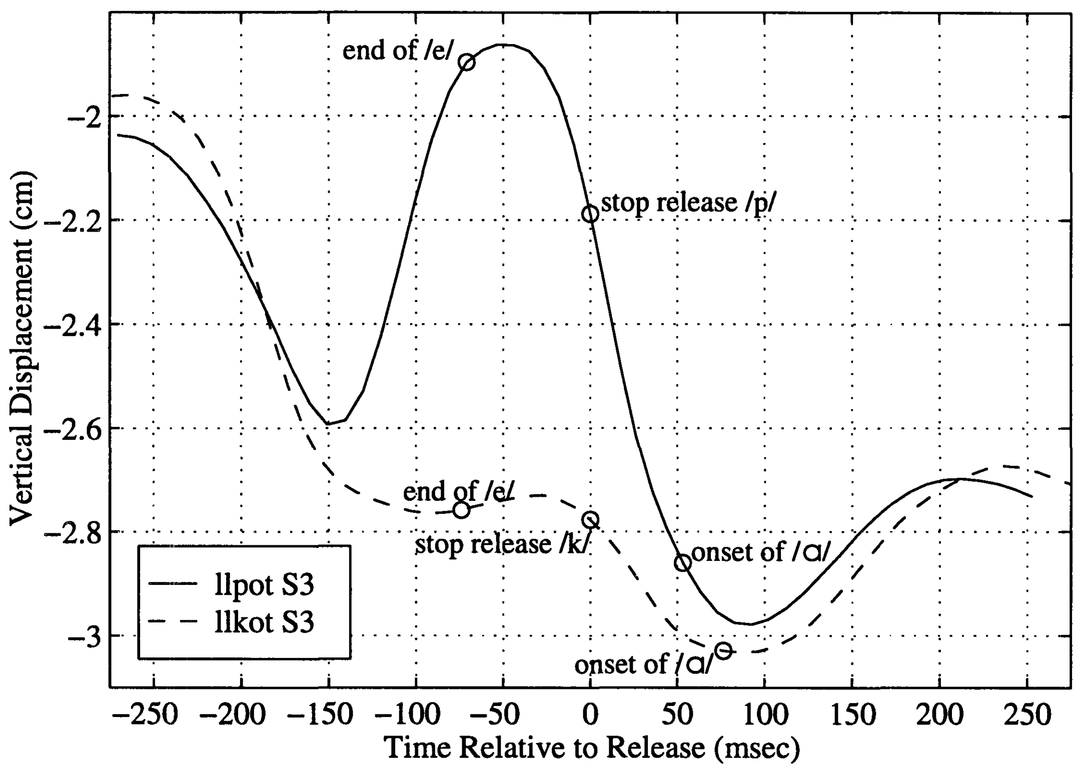

Figure 3-20: Average vertical displacement of the lower lip, from a fixed reference point in the vocal tract, as a function of time relative to the /p/ and /k/ stop releases in the utterances pot and kot, respectively, spoken by Subject 3. Refer to Section 3.3.1 for a description of the averaging technique. The important acoustic events are labeled.

3.4 Velocity Analysis

The maximum downward velocity occurring closest to the time of the stop or nasal release was measured for several articulators. The maximum downward velocity was selected as a readily available measure of production of the given sound. The first section describes the procedure used to determine the maximum velocity, and the second section includes several tables of maximum velocities, as well as discussion of several trends observed in the velocities across speakers and utterances.

3.4.1 Procedure

To determine the maximum downward velocity for a given articulator and utterance, a window (nominal duration 60 msec) was centered at the time of the release in each velocity waveform repetition (calculated by taking the derivative of the corresponding transducer displacement waveform repetition), and the maximum downward velocity was chosen from within that window. Then, the times and magnitudes of the maxima were averaged together, respectively, resulting in a single, average maximum and its corresponding time relative to the release. Occasionally, as reflected in the tables of Section 3.4.2, the window duration had to be increased to greater than 60 msec to capture the maximum for an alveolar or velar stop consonant with a maximum downward velocity occurring later than 30 msec following the release.

3.4.2 Results and Discussion

The maximum downward velocities for the lower lip, tongue blade, tongue body and lower jaw articulators are in Tables 3.1, 3.2, 3.3 and 3.4, respectively. In Section 3.3.2 various velocity values were discussed in the context of observations made from the displacement waveforms. Those values are included in the tables in the present section. This section is devoted to trends able to be observed from the velocity tables.

The maximum downward velocities for the lower lip are shown Table 3.1 for utterances containing a labial stop or nasal consonant. The first observation is that the lower lip velocities are slower for utterances containing the vowel /i/ than those containing the vowel /ɑ/. The lower lip rides on the lower jaw, so the difference in rates of lower lip movement could be attributed to the difference in final jaw heights (and, consequently, rates of lower jaw movement) required for the production of the following vowel, since a lower jaw position is required for the vowel /ɑ/ than the vowel /i/. A second observation is that Subject 2 repeatedly has considerably more rapid lower lip velocities than either of the other two subjects, indicating a consistent difference in speaker strategies.

Utterance	Subject 1		Subject 2		Subject 3	
	Time msec	Vel cm/s	Time msec	Vel cm/s	Time msec	Vel cm/s
peet	3.4	14.8	27.1	21.0	13.0	11.6
pot	17.3	17.6	28.2	22.6	18.1	16.5
speet	13.2	12.5	17.9	20.9	15.3	9.5
spot	20.2	15.3	25.8	26.1	26.1	13.8
meet	-1.7	13.9	27.9	22.3	18.9	12.5
mot	15.8	23.7	19.4	26.8	14.9	18.9
smeet	4.1	14.3	10.4	24.8	9.8	12.4
smot	16.5	19.0	18.5	28.5	8.6	17.4
beet	-	-	-	-	13.1	11.7
bot	-	-	-	-	12.4	17.4

Table 3.1: Lower lip average maximum downward velocities occurring closest to the time of the labial stop or nasal release in selected utterances. Corresponding average times are given relative to the release. A negative time indicates that the average maximum velocity occurred prior to the release. Utterances containing the voiced labial stop were not recorded by Subjects 1 and 2. Refer to the text for method of measuring the velocity.

The maximum downward velocities for the tongue blade are displayed in Table 3.2 for utterances containing the alveolar stop and nasal consonants. For utterances containing the velar stops, the tongue body maximum downward velocities are in Table 3.3. The findings from this data are less conclusive, since the tongue is very flexible and pliant, thus movements recorded at the transducer locations are not highly indicative of the movements at nearby constriction locations. Once again, the velocities for utterances containing the vowel /i/ are generally smaller than for utterances containing the vowel /ɑ/. This observation is particularly true for the tongue body velocities, since the tongue body moves very little to produce the sequences /ki/ and /gi/, in comparison to the movement for the sequences /kɑ/ and /gɑ/ (Subject 3). For example, the tongue body velocities for the utterances kot and got are very similar, 16.1 cm/sec and 17.4 cm/sec, respectively. The velocities for the utterances keet and geet, 2.0 cm/sec and 2.7 cm/sec, respectively, are on the order of 15 cm/sec smaller. The tongue body maximum velocities occur well after the stop release. The tongue body appears to be reaching its peak velocity as it moves into position for the vowel steady-state, as opposed to sometime during the production of the stop.

Utterance	Subject 1		Subject 3	
	Time msec	Vel cm/s	Time msec	Vel cm/s
teet	-4.9	1.6	21.4	4.4
tot	51.7	13.7	34.3	22.8
steet	21.7	-0.2	21.0	2.2
stot	48.7	12.1	32.8	15.8
neet	-	-	6.1	4.4
not	-	-	15.8	22.9
sneet	-	-	13.7	3.9
snot	-	-	12.1	14.5
deet	-	-	11.9	4.3
dot	-	-	29.4	24.3

Table 3.2: Tongue blade average maximum downward velocities occurring closest to the time of the initial alveolar stop or nasal release in selected utterances. Corresponding average times are given relative to the release. A negative time indicates that the average maximum velocity occurred prior to the release. A negative velocity indicates an articulator is moving in the upward (not downward) direction. Data was unavailable for the tongue blade transducer for Subject 2 and utterances containing the alveolar voiced stop or nasal were not recorded by Subject 1. Refer to the text for method of measuring the velocity.

The final table, Table 3.4, contains the lower jaw maximum downward velocities for all recorded utterances. Many of the observations regarding the data in this table have already been made in Section 3.3.2. Additional observations include: (1) The velocities for the voiced stops are very similar to those of the voiceless stops, within a given vowel. Based on data from only one subject, the voiced stops have a consistently higher velocity for the vowel /ɑ/ than the voiceless stops; (2) All utterances (not just those discussed in Section 3.3.2) exhibit the trend of a faster

Utterance	Subject 1		Subject 2		Subject 3	
	Time msec	Vel cm/s	Time msec	Vel cm/s	Time msec	Vel cm/s
keet	27.2	5.1	51.2	6.4	29.7	2.0
kot	62.0	12.6	27.7	14.6	49.9	16.1
skeet	16.6	4.0	41.0	5.9	44.2	0.7
skot	56.4	13.9	23.5	15.9	50.3	15.9
geet	-	-	-	-	48.1	2.7
got	-	-	-	-	51.5	17.4

Table 3.3: Tongue body average maximum downward velocities occurring closest to the time of the velar stop release in selected utterances. Corresponding average times are given relative to the release. Utterances containing the voiced velar stop were not recorded by Subjects 1 and 2. Refer to the text for method of measuring the velocity.

velocity when they contain the vowel /ɑ/ (vs. /i/); (3) The velocities for utterances containing /n/ are consistently less than (or in one case, equal to) the corresponding utterances containing /m/. For example, the velocity for meet is 1.1 cm/sec whereas the velocity for neet is 0.6 cm/sec (Subject 3); and (4) In general, the maximum downward velocity is reached sooner following the release in utterances containing /i/ than /ɑ/.

One observation in particular can be made upon examination of two or more of the articulators' velocities. When the tongue body is the primary articulator, the maximum velocity is generally greater for the tongue body than the lower jaw, agreeing with the observation made in Section 3.3.2 that the rate of the primary articulator (in that section, the lower lip) is greater than that of the secondary articulator (the lower jaw). For the tongue blade, however, the velocities are much more similar, attesting to the fact that the lower jaw is involved to a greater degree in the formation of the actual constriction and release. The lower jaw is required to move upward to facilitate the contact of the tongue tip with the palate during the formation of the constriction.

The average maximum jaw velocities were greatest for the utterances containing the alveolar stop /t/, ranging up to slightly above 11 cm/sec. This value is somewhat larger than the range of values found by Smith and Gartenberg (1984) and Smith and McLean-Muse (1986) of 6 - 8.5 cm/sec in speech. The 11 cm/sec value does fall

Utterance	Subject 1		Subject 2		Subject 3	
	Time msec	Vel cm/s	Time msec	Vel cm/s	Time msec	Vel cm/s
peet	-4.9	4.1	7.1	1.9	2.7	1.2
pot	17.1	6.8	22.0	4.2	10.8	6.3
speet	1.7	4.7	-7.1	2.6	-5.2	2.1
spot	9.5	8.0	17.0	5.8	13.6	7.7
teet	17.8	2.7	23.7	2.4	29.8	1.3
tot	57.3	11.1	38.1	6.5	37.4	6.9
steet	20.7	2.5	10.3	2.2	14.6	2.7
stot	52.1	10.9	30.3	8.0	26.5	8.6
keet	12.2	0.6	43.7	0.9	-19.9	0.5
kot	55.0	4.8	35.2	2.9	30.4	4.2
skeet	-10.3	4.5	-4.1	1.2	-22.2	3.3
skot	21.0	9.0	12.3	5.2	17.4	9.8
meet	-12.4	3.1	12.9	1.6	-1.3	1.1
mot	11.5	9.3	9.4	5.1	2.0	5.5
smeet	-5.9	5.0	-12.2	3.0	-4.1	2.7
smot	10.2	8.8	-2.7	6.6	-6.4	7.8
neet	-	-	-	-	-4.8	0.6
not	-	-	-	-	13.5	5.5
sneet	-	-	-	-	-4.5	2.1
snot	-	-	-	-	4.1	7.0
beet	-	-	-	-	-3.7	1.2
bot	-	-	-	-	4.5	6.7
deet	-	-	-	-	21.4	1.1
dot	-	-	-	-	32.3	7.9
geet	-	-	-	-	-20.0	0.6
got	-	-	-	-	35.8	6.2

Table 3.4: Lower jaw average maximum downward velocities occurring closest to the time of the stop or nasal release in all recorded utterances. Corresponding average times are given relative to the release. A negative time indicates that the average maximum velocity occurred prior to the release. Utterances containing the voiced stops or alveolar nasal were not recorded by Subjects 1 and 2. Refer to the text for method of measuring the velocity.

well within the 10 - 20 cm/sec range given by Nelson et al., 1984, Kent and Moll, 1972, and Beckman and Edwards, 1993, for peak velocities of mandible movement. The average maximum lower lip velocities extended up to the range 27 - 28.5 cm/sec, slightly higher than estimates from Fujimura, 1961b, and Sussman et al., 1973, of 25 cm/sec or less for the peak velocities of movement of the body of the lower lip following release of a labial stop consonant. For the tongue tip, average maximum velocities were recorded up to about 20 - 25 cm/sec in the present study, which is at the low end of the 20 - 40 cm/sec range reported by Kent and Moll, 1972. Finally, average peak velocities for the tongue body were measured to be in the range of 15 - 17 cm/sec, slightly less than the 20 cm/sec value determined by Kent and Moll, 1972, but at the high end of the 10 - 15 cm/sec range reported by Kuehn and Moll, 1976.

3.5 Summary

In Section 3.1 the speakers' backgrounds and the selected corpus were described. In Section 3.2 the operation of the recording device, an electro-magnetic midsagittal articulometer (EMMA) was outlined.

The first part of Section 3.3 contains a description of the data processing procedure developed to average the displacement data. Through the use of linear time warping, the averaging method accounts for the presence of slight variations in speaking rate between repetitions to yield a representative sample of the displacement waveform for each utterance. The second part of Section 3.3 is the analysis of the averaged displacement waveforms. Listing a few of the observations made from the data: (1) The average rate of maximum downward jaw movement near the time of the stop release is greater for utterances in which the stop is preceded by /s/. The jaw is constrained to maintain a high position until the end of /s/ production. The jaw then moves rapidly downward to compensate for the high position and allow production of the following vowel to occur in a timely fashion; (2) When a stop or nasal consonant is followed by the vowel /ɑ/, the jaw moves downward more rapidly and has a greater relative displacement than when followed by /i/. This observation provides further evidence of the correlation between larger displacements and faster rates of movement for the articulators; (3) For the production of /p/ in pot, Subject 2 has a strategy of making a larger lower lip movement, with a more rapid maximum velocity following the release, than the other two subjects. Subject 2 also takes more time than the other speakers to produce the stop, possibly indicating incomplete compensation for the larger movements; (4) A bandwidth of about 3 - 5 Hz exists for cyclic articulatory movements produced in speech, such as the movement of the lower jaw from the steady-state portion of the vowel /e/ in say to the steady-state portion of the vowel /ɑ/ in pot, spoken in the phrase, "Say pot again."; and (5) The duration of time a primary articulator takes to produce a stop (including the onset phase, closure interval and offset phase) is about 200 - 300 msec, equivalent to the length of one period of the cyclic movement.

In Section 3.4, the average maximum downward velocities of several articulators following the release were examined. In addition to providing further basis for the observations made in Section 3.3, the maximum velocities were found to compare favorably to the maximum velocity values in the literature.

In summary, several observations were made in Chapter 3 about the displacement and velocity of the articulators during stop and nasal production. The observations were related to coarticulation, varying speaker strategies, bandwidth limitations on articulator movements, and average maximum velocities recorded by the transducers. The research may contribute to theories of control and coordination of the articulatory structures during speech production. In Chapter 4, the simultaneously-recorded acoustic data will be analyzed. In addition, the articulation and acoustic data will be utilized in association with the models of Chapter 2 to develop a better understanding of certain aspects of stop-consonant production.

Chapter 4

Acoustic Analysis and Modeling

The methods and results of the acoustic analysis are presented in this chapter. The incorporation of the acoustic and articulation data into the models discussed in Chapter 2 will also be presented. The first section provides a description of the speakers' backgrounds and the corpus. The second section discusses the acoustic recording method utilized in the study. The third section describes the measurement of the frication noise burst in the acoustic data. The fourth section involves derivation of the rate of constriction cross-sectional area increase following the stop-consonant release. The rate of area increase is derived from the articulator movements, with the aid of the low-frequency circuit model introduced in Chapter 2, Section 2.1, and the noise burst duration as measured from the acoustic data in the third section of the present chapter. The fifth section involves supplementation of the initial portion of the $F1$ transition following the stop release, measured from the acoustic data, with the $F1$ transition calculated from the constriction cross-sectional area and the high-frequency circuit model of Chapter 2, Section 2.2. The final section summarizes the results of the acoustic analysis and the refinements realized in the models.

4.1 Speakers and Corpus

The speakers and corpus for the acoustic experiment are identical to the three speakers and the corpus of the articulation experiment, since the experiments were performed simultaneously (refer to Chapter 3, Section 3.1).

4.2 Recording Method

The amplified output of the acoustic speech signal was recorded by a directional microphone during each of the three electro-magnetic midsagittal articulometer (EMMA) system experiments. The acoustic signal, digitized at a sampling rate of 10 kHz, was part of a single, digitized signal that included both the articulation and acoustic data recorded during the experiment. Signal processing software (Henke, 1989; Perkell et al., 1992) was used to demultiplex the digitized signal into separate, synchronized acoustic and articulatory signal streams.

4.3 Frication Noise Burst Determination

Three different types of noise are generated following the stop consonant release. Sequentially, the types of noise are the transient burst, the frication noise burst and the aspiration noise. This section will be devoted to determining the duration and general shape of the frication noise burst. Emphasis will be placed on the unaspirated /p/ and /t/, spoken in the utterances <u>spot</u>, <u>speet</u>, <u>stot</u> and <u>steet</u>, for which the frication noise burst is readily detectable via the procedure described in this section.

4.3.1 Procedure

The signal processing software KLSPEC, developed by D. H. Klatt, was utilized to measure the frication noise burst in the acoustic data. A very short (3.2 msec) Hamming window was centered at the time of the stop-consonant release for each repetition of a particular utterance for a given speaker. The stop release time was rounded to the nearest millisecond due to KLSPEC software restrictions. A short window duration was selected to observe minute changes in the noise amplitude with time. The 512-point discrete Fourier transform was computed for the first-differenced waveform multiplied by the Hamming window. Averaging of the DFT magnitude spectra was performed across all repetitions of the given utterance and speaker at the stop release time. Then, the window was moved to one msec following the release in each repetition and averaging was performed again. The procedure continued, in one-msec increments, until well after the termination of the frication noise burst for the unaspirated /p/ or /t/, approximately 15 - 20 msec after the release.

The frication noise burst is generated near the constriction, throughout the time interval when the constriction is sufficiently narrow. The turbulence noise excites the cavity in front of the constriction. For the production of /p/, in which there is no front cavity, there are no resonances to excite and the resultant spectrum is smooth with no major prominences. The peak amplitude (in dB) was measured within the 2 - 3 kHz frequency band in each averaged spectrum of spot and speet. The peak amplitude in this frequency range is believed to be representative of the peak amplitude of the overall averaged spectrum. For the production of /t/, a short front cavity, approximately 2 cm in length, exists in front of the constriction. The frication noise burst excites the resonances of this front cavity, the lowest resonance of which is within the 3 - 5 kHz range. In most cases, the peak amplitude (in dB) within the 3 - 5 kHz frequency band was measured in each averaged spectrum of stot and steet. There were a few instances when the major spectral prominence and the lowest excited front cavity resonance appeared to be at different frequencies. On these occasions, the peak amplitude of the prominence that was surmised to be the front cavity resonance was measured instead. The author's judgment was required to make the distinction. The peak noise amplitude determined in this fashion from the averaged spectra was plotted as a function of time following the stop release for each utterance and each speaker.

The time course of the noise generated following the stop release was examined to determine the duration of the frication noise burst. In general, the duration of the burst was measured as the time interval during which the amplitude of the noise continually remained within 10 dB of the maximum noise amplitude following the release. If there was no apparent burst onset, i.e., no point within a few milliseconds after release when the amplitude of the noise burst was at least 10 dB below the maximum burst amplitude, the initiation of the frication noise burst was chosen to be at the time of the stop release in the averaged spectrum. For the utterances containing an unaspirated /p/, if all the points in the time interval 6 - 11 msec following the stop release had amplitudes above the 10 dB down point, the end of the noise burst was chosen to be the point in time within that time interval when the amplitude was at a global minimum. For the utterances containing an unaspirated /t/, the time interval in which to locate the global minimum increased to 11 - 16 msec. It was observed that utterances containing /i/ often had several local minima within the specified ranges, making it less certain that the correct choice for the termination of the frication noise should be the global minimum within the range.

In addition to measuring the duration of the frication noise burst from the averaged spectrum, two other durations of interest were measured. For each utterance and speaker, the average time from the stop release to the onset of the noise burst was measured. The average time from the end of the noise burst to the onset of voicing for the following vowel was also measured. The voice onset time (VOT) was calculated to be the sum of these three durations.

4.3.2 Results and Discussion

The average time course of the noise generated following the stop release is shown in Figures 4-1 and 4-2 for the utterances spot and stot, respectively, spoken by Subject 3. Table 4.1 lists the measured average durations, rounded to the nearest tenth millisecond, of the noise bursts for each of the unaspirated utterances for each speaker. The table also gives the average duration from the time of the stop release to the onset of the frication noise burst, the average duration from the end of the frication burst to the onset of voicing of the following vowel, and the voice onset time for each of the utterances and speakers. (The acoustic data from Subject 2 will not be analyzed for utterances containing /t/ due to the fact that, as discussed in Chapter 3, Section 3.2, the tongue front (TF) transducer was not functioning properly during the experiment, resulting in the inability to perform the remaining analysis in Sections 4.4 and 4.5 of the present chapter.)

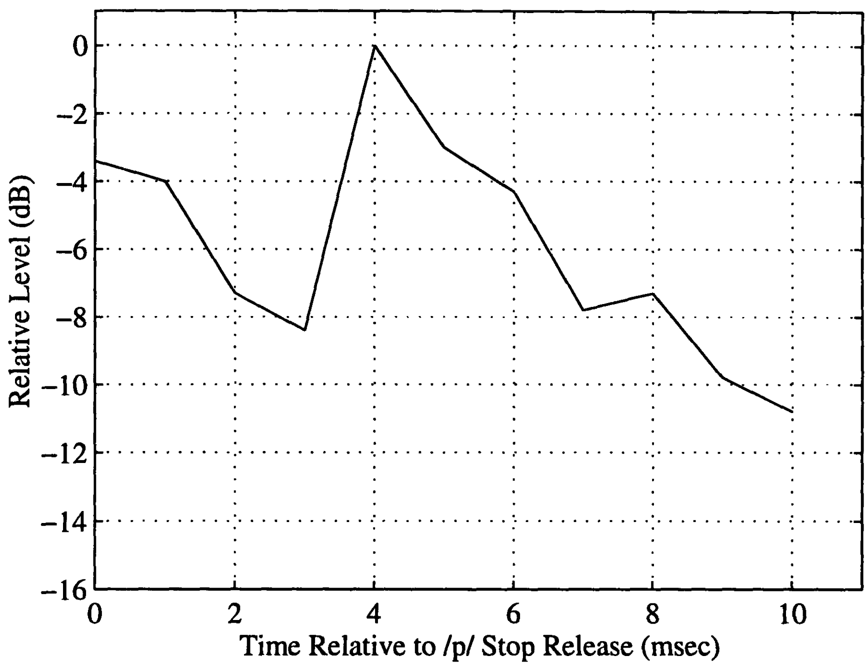

Figure 4-1: Time course of the noise generated following the /p/ stop release in spot, spoken by Subject 3. The amplitude of the noise is relative to the maximum noise amplitude, occurring 4 msec following release. In this example, the frication noise burst begins at the time of the stop release and extends for a duration of 9.2 msec. See text for method of measuring the burst.

In Figure 4-1, the effects of the production of the transient burst are observed to carry over into the production of the frication noise burst. The transient burst occurs during the first 0.5 - 1 msec following the stop release. Occasionally, in some utterances, the production of the transient noise influences the first few milliseconds of the production of the frication noise burst, resulting in the occurrence of a local minimum in the noise amplitude up to a few milliseconds following the release. A similar effect is seen for the utterance spot spoken by Subject 1 (not shown). The frication noise burst produced by Subject 2 for the utterance spot (not shown) was not influenced by the transient, however. Although the frication burst for the utterance stot spoken by Subject 1 (not shown) reveals a large transient burst effect, the same utterance spoken by Subject 3 (Figure 4-2) is only minimally, if at all, affected by the transient.

A few observations can also be made about frication noise burst production from Table 4.1. Comparing across subjects, each speaker has a frication noise burst du-

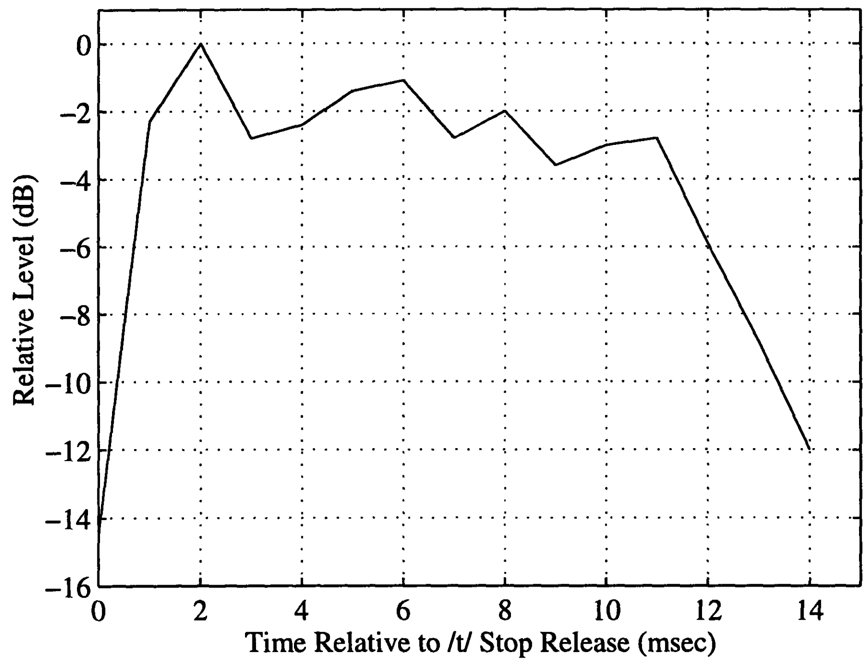

Figure 4-2: Time course of the noise generated following the /t/ stop release in <u>stot</u>, spoken by Subject 3. The amplitude of the noise is relative to the maximum noise amplitude, occurring 2 msec following release. In this example, the frication noise burst begins 0.4 msec following the time of the stop release and extends for a duration of 12.9 msec. See text for method of measuring the burst.

ration in the range 7 - 10 msec for utterances containing the unaspirated /p/. The duration of the frication noise burst for utterances containing the unaspirated /t/ is in the range 11 - 16 msec. Within subjects, as well as across subjects, the frication noise burst following the release of an unaspirated /t/ is longer than the burst following the unaspirated /p/ release. The onset of the noise burst is at or very near the stop release in each utterance, for each speaker. The amount of time between the end of the frication burst and the onset of the following vowel varies widely, however, both between utterances and between subjects. For example, the duration between the end of the noise burst and the onset of the vowel is much longer (16 - 17 msec) for <u>stot</u> spoken by Subject 1 than spoken by Subject 3 (2 - 7 msec).

	Subject 1				Subject 2				Subject 3			
	T1	NB	T2	VOT	T1	NB	T2	VOT	T1	NB	T2	VOT
Utt.	ms	ms	ms	ms	ms	ms	ms	ms	ms	ms	ms	ms
spot	0.0	8.0	3.7	11.7	0.0	9.0	9.8	18.8	0.0	9.2	7.3	16.5
stot	0.0	14.0	16.0	30.0	-	-	-	-	0.4	12.9	2.2	15.5
speet	0.0	7.0	12.1	19.1	0.4	8.6	11.8	20.8	0.0	10.0	9.8	19.8
steet	0.1	11.2	16.8	28.1	-	-	-	-	0.4	15.6	6.3	22.3

Table 4.1: Duration from time of stop release to onset of noise burst (T1), frication noise burst duration (NB), duration from end of noise burst to voice onset (T2) and voice onset time (VOT) are given for each of the unaspirated utterances of each speaker. The VOT is obtained by adding all three durations together (T1 + NB + T2 = VOT). All durations are rounded to the nearest tenth of a millisecond. (For utterances containing /t/, the acoustic data of Subject 2 was not analyzed. See text for explanation.)

4.4 Constriction Cross-sectional Area Derivation

A method for estimating a linear rate of constriction cross-sectional area increase following the stop-consonant release has been developed for each of the labial and alveolar constriction locations. The method incorporates both acoustic and articulation data. An initial value for the linear rate of area increase, derived from the movements of one or two transducers, will be used as an input to the low-frequency circuit model introduced in Chapter 2, Section 2.1. One of the model outputs, the airflow through the constriction, will be utilized in conjunction with the initial linear area rate in a second model, estimating the frication noise sound source. The duration of the noise burst calculated from the second model will be compared to the noise burst duration from the acoustic data, measured in the previous section. The linear rate of area increase which results in a modeled noise burst duration equal to the measured noise burst duration is the rate chosen for a given utterance and speaker.

4.4.1 Method

Deriving an estimate of the linear rate of constriction cross-sectional area increase following the stop-consonant release first involves constructing a model of the cross-sectional area. The model should utilize articulator displacements measured by the

EMMA system. For the stop consonant /p/, one possible model is illustrated by Figure 4-3, in which the lip opening is modeled by an ellipse with cross-sectional area $A_p(t)$.

$$A_p(t) = k\frac{\pi}{4}v(t)h(t) \tag{4.1}$$

Each of the lip transducer displacements was averaged across repetitions for a given

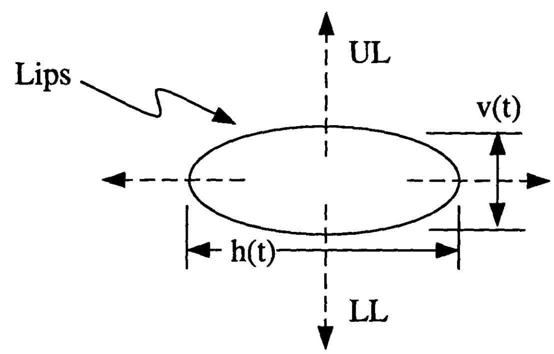

Figure 4-3: Labial constriction cross-sectional area ellipse model. The vertical lip separation, $v(t)$, is derived from the upper lip (UL) and lower lip (LL) transducer displacements (vertical dashed lines). The horizontal lip separation, $h(t)$, is based on experimental work by Fujimura(1961b) (horizontal dashed lines).

utterance and speaker, as described in Chapter 3, Section 3.3.1. The vertical lip separation, $v(t)$, is the difference between the upper lip (UL) and lower lip (LL) average displacements, after removing the dc offset caused by non-zero transducer separation at the time of the release. As the lips open following the release, the distance between the UL and LL transducers increases, resulting in an increasing $v(t)$. The horizontal separation of the lips, $h(t)$, cannot be measured by the EMMA system since the system is only capable of recording movements within the midsagittal plane, and the horizontal lip movements are perpendicular to that plane. The horizontal lip separation was modeled as being proportional to $(1 - e^{-t/\tau})$, where $\tau = 3$ msec, based on experimental work by Fujimura (1961b). Lastly, in order to avoid misrepresenting the *rate* of area increase as zero right at the time of the release, the starting value of the horizontal lip separation, $h(0)$, was arbitrarily set to 0.5 cm.

As noted above, the UL and LL transducers are not located at the edges of the lip constriction. Instead, in order to prevent them from interfering with production, the transducers were placed on the vermillion borders of the lips. There is believed to be considerable difference in movement between the inner edges of the lips and the vermillion borders. The release of intraoral pressure at the time of the stop release is thought to cause the inner edges of the lips to be blown rapidly apart. Furthermore, an ellipse is only a rough approximation of the shape of the lip opening. To partially compensate for the discrepancy in movement between the actual labial constriction opening and the model, the area of the ellipse was multiplied by a constant. Considering only the effects of the intraoral pressure release, the value of the constant would be expected to be greater than 1.0. It should be noted that, even with the constant multiplier, the model probably still does not accurately reflect the movement of the lips during the first millisecond or so following the release, when the inner lip edges are moving the most rapidly apart. The release of intraoral pressure has been observed to establish a damped oscillation of the lips (Fujimura, 1961b). The oscillation is not seen in the movements recorded by the transducers, and consequently is not represented in the model. One effect of the lip oscillations may be to slow the rate of area increase during a time interval starting a few milliseconds after the release (corresponding to the first inward lip vibratory movement). This effect on the area is not incorporated in the model.

For the stop consonant /t/, one possible constriction cross-sectional area model appears in Figure 4-4. The constriction area formed by the hard palate and the

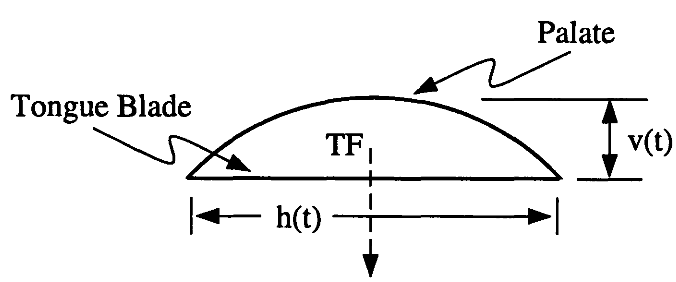

Figure 4-4: Alveolar constriction cross-sectional area segment of a circle model. The vertical separation between the hard palate and the tongue tip, $v(t)$, is derived from the tongue front (TF) transducer displacement (vertical dashed line). The horizontal effective tongue width, $h(t)$, is determined from $v(t)$ and the radius of the circle in the segment model.

tongue blade was modeled as the segment of a circle having a fixed radius R. To determine R, measurements of the palate height and width were made from dental casts of each subject. The constriction location was estimated to be aligned with the distal surfaces of the upper canines. Palate height, v_o, was defined to be the vertical distance in the midsagittal plane from the point where the tongue touches the palate to the occlusal surface of the upper teeth. Palate width, h_o, was defined to be the distance between the lingual surfaces of the upper canine teeth, measured just below the occlusal surface. For Subject 1, $v_o = 0.95$ cm and $h_o = 2.6$ cm, and for Subject 3, $v_o = 1.15$ cm and $h_o = 2.7$ cm. The constriction cross-sectional area was not determined for Subject 2, for whom the tongue front (TF) transducer data was unavailable. The movements of the tongue front or tip, as recorded by the TF transducer, were averaged across repetitions for a given utterance and speaker, as described in Chapter 3, Section 3.3.1. The downward movement of the TF transducer from the palate to the position for the following vowel is the increasing height, $v(t)$, of the segment of the circle. Similar to the model for /p/, the segment width, $h(t)$, cannot be obtained using movement recorded by the EMMA system. Consequently, $h(t)$ is determined from $v(t)$ and the constant radius R. To avoid misrepresenting the rate of area increase as zero right at the time of the release, the initial width, $h(0)$, was set to 0.1 cm for utterances containing the vowel /ɑ/ and 0.3 cm for /i/. Although the values for $h(0)$ were arbitrarily chosen, a larger value was selected for /i/ because the tongue is believed to have more palatal contact at the time of the release for utterances containing /i/, in anticipation of the tongue position required for the upcoming vowel. A dc offset, representing the difference in lip transducer locations at time zero, was subtracted from $v(t)$, resulting in $v(0) = 0$. The set of equations determining the segment area are shown below, where $\theta(t)$ is the central angle of the sector of the circle, and $A_t(t)$ is the estimated constriction cross-sectional area.

$$R = \frac{1}{2v_o}\left(v_o^2 + \frac{h_o^2}{4}\right) \tag{4.2}$$

$$h(t) = 2\sqrt{2kv(t)R - kv(t)^2} + h(0) \tag{4.3}$$

$$\theta(t) = 2sin^{-1}\left(\frac{h(t)}{2R}\right) \tag{4.4}$$

$$A_t(t) = \frac{\theta(t)}{2}R^2 - \frac{h(t)}{2}(R - kv(t)) \tag{4.5}$$

Similar to the UL and LL transducers, the TF transducer was not located at the actual point on the tongue tip that forms the constriction. Instead, in order to prevent the transducer from interfering with production, it was placed approximately 3 - 4 mm behind the tip of the tongue. Because the tongue is flexible and pliant, there is considerable difference in movement between the tongue tip and a point a few millimeters behind the tip. In addition, the segment of a circle is only a rough approximation of the shape of the alveolar constriction. The segment approximation assumes that the tongue remains flat as it moves away from the closure, which is unlikely. To partially compensate for the discrepancy in movement between the actual alveolar constriction opening and the model, the height of the segment of the circle, $v(t)$, was multiplied by a constant k prior to calculating the segment width and area. The constant multiplier is anticipated to have a value greater than 1.0 since, near the release, the tongue tip is expected to move more rapidly than a point behind the constriction. In addition, it should be noted that, even with the constant multiplier, the segment area model is still thought not to accurately reflect the movement of the tongue tip during the first millisecond or so following the release. During that time interval, the pressure behind the constriction is believed to cause a downward force that accelerates the tongue tip movement. This acceleration was not measured by the transducers, and consequently was not represented in the model.

The labial and alveolar constriction models were observed to yield a cross-sectional area rate of increase which was too slow during the initial 10 - 20 msec following the release, and too rapid for the next 30 - 40 msec, compared to estimates based on the acoustic data. It was decided to linearize the rate of area increase over the first 10 msec following the release, and assign that linear rate to be the cross-sectional area rate of increase for the first 50 msec following the release. The time interval of interest in the constriction cross-sectional area is the time from the stop release to about 20 msec after the release, encompassing the frication noise burst. The 10-msec time interval immediately following the release was chosen for the linearization because it is roughly the average of the measured frication noise burst durations of Section 4.3. Although the linear rate may not be considered to be a good representation of the change in cross-sectional area as time progresses following the release, particularly once vowel steady-state has been reached, the study is interested in no more than the first 20 msec after the release.

Even after the linearization, the initial rate during the first 10 - 20 msec following the release remained too slow. A refinement process was devised to determine a linear rate which more accurately reflected the rate of movement following the release. This process involved adjusting a constant multiplier, k, contained in each of the constriction area formulas. (See Equations 4.1 and 4.5.) A description of the process follows. The first step in refining the linear rate of area increase was to generate and linearize an initial estimate of the area rate of increase from the constriction models, as described above. In the second step, the linear rate of area increase was utilized as the parameter A_c in the low-frequency circuit model of Chapter 2, Section 2.1. The model generates the airflow through the constriction, U_c, which in turn is an input to a second model, estimating the time course of the frication noise burst following the release. The duration of the frication noise burst was measured using the same method as in Section 4.3. In the third step, the duration of the noise burst derived from the articulation data (via the employment of models) was compared to the noise burst duration measured from the acoustic data (refer to Section 4.3). Depending upon whether the modeled noise burst duration was longer or shorter than the actual burst, the value of k was adjusted and the linear rate recalculated. This procedure was repeated, starting with step 2, until the two burst durations were in agreement to within less than a millisecond. Linear rates following release of the unaspirated stop consonants /p, t/ were determined for each utterance and speaker via this process.

4.4.2 Results and Discussion

A set of graphs representing the circuit and frication noise burst model inputs and outputs is shown in Figure 4-5 for the utterance spot, spoken by Subject 3. Comparing the duration of the noise burst shown in Figure 4-5 (d) with the noise burst duration from the acoustic data, Figure 4-1 and Table 4.1, it can be seen that the durations are both 9.2 msec. The linearized constriction cross-sectional area is shown in Figure 4-5 (a). The linear area, enlarged, is redisplayed in Figure 4-6, with its corresponding area generated by the labial constriction model. The linear rate of area increase was determined to be 39.0 cm^2/sec.

Similarly, for the utterance stot, spoken by Subject 3, the set of graphs representing the circuit and frication noise burst model inputs and outputs is shown in Figure 4-7. Comparing the duration of the noise burst shown in Figure 4-7 (d) with the noise burst duration from the acoustic data, Figure 4-2 and Table 4.1, it can be seen that the durations are both 12.9 msec. The resulting constriction cross-sectional area is shown in Figure 4-7 (a). The linear area is redisplayed in Figure 4-8, with its corresponding area generated by the alveolar constriction model. The linear rate of area increase was determined to be 25.5 cm^2/sec.

The example described in Chapter 2, Section 2.1, contains the set of graphs representing the circuit and frication noise burst model inputs and outputs in Figure 2-3 for the unaspirated /p/ in the utterance spot, spoken by Subject 1. Comparing the duration of the noise burst shown in Figure 2-3 (d) with the noise burst duration from the acoustic data, Table 4.1, both durations are equal to 8.0 msec. The resulting constriction cross-sectional area is shown in Figure 2-3 (a). The linear area is enlarged and redisplayed in Figure 2-5, with its corresponding area generated by the labial constriction model. The linear rate of area increase was determined to be 46.6 cm^2/sec.

A constant multiplier was present in each of the formulas for the labial and alveolar constriction cross-sectional area models. (Refer to Equations 4.1 and 4.5.) The value of the constant for the labial stop ranged from 1.6 - 3.6, depending upon the speaker. The constant compensated for the difference in rates of movement of the

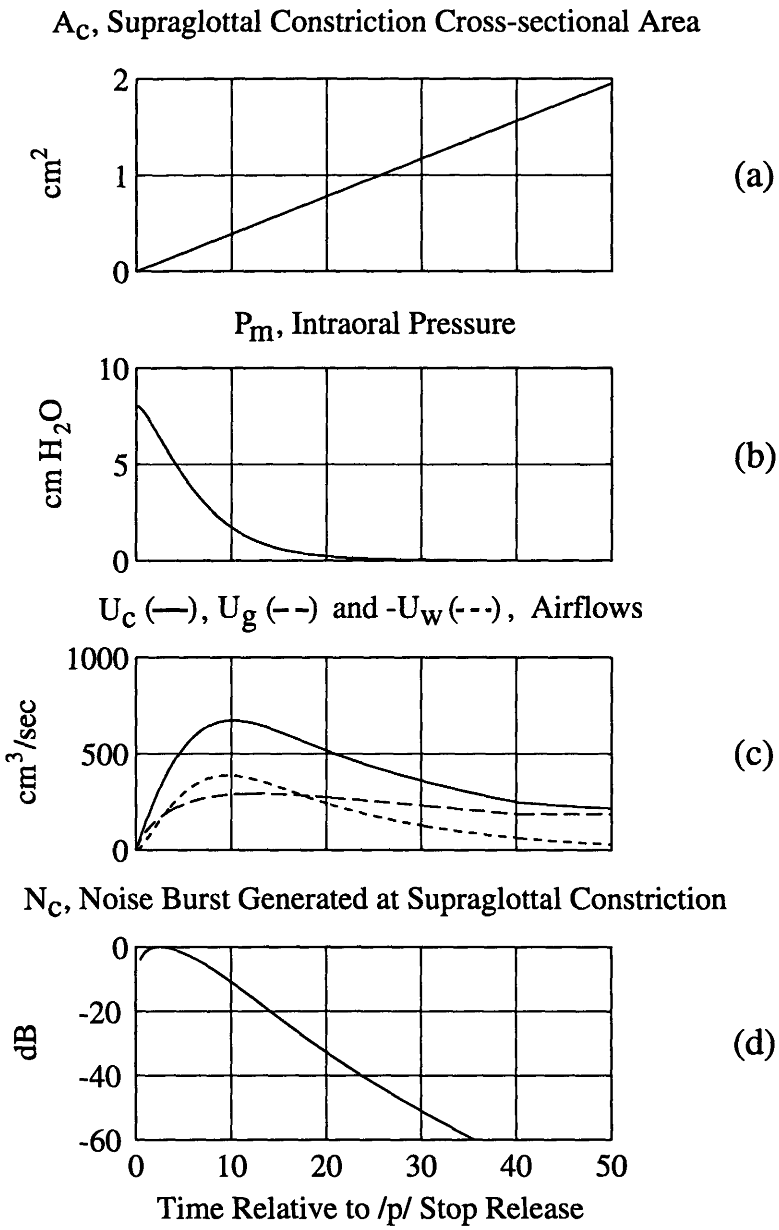

Figure 4-5: Low-frequency circuit and frication noise burst model inputs and outputs for the unaspirated stop consonant /p/ upon release of closure in the utterance <u>spot</u>, based on data derived from Subject 3: (a) Lip-opening constriction cross-sectional area, A_c; (b) Pressure within the mouth, P_m; (c) Airflow through the lip-opening constriction, U_c (solid line), airflow through the glottis, U_g (dashed line), and airflow generated by the inward displacement of the vocal-tract walls, -U_w (dotted line) (the negative sign indicates the direction of displacement of U_w is inward); (d) Frication noise burst, N_c. Time zero is the instant of stop release.

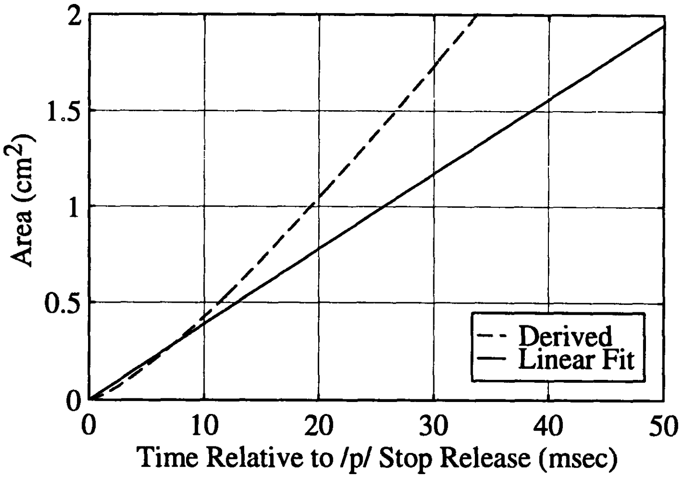

Figure 4-6: Labial constriction cross-sectional area for the utterance spot, based on data derived from Subject 3. The dashed line is the area derived from the articulation data, models, and acoustic data, as discussed in the text. The solid line is the best linear fit to the first 10 msec of the derived area. The linear rate of area increase is 39.0 cm^2/sec.

vermillion borders and the inner edges of the lips. The inner edges move more rapidly, particularly near the time of the release. The value of the constant multiplier for the alveolar stop ranged from 2.1 - 4.3 for utterances containing the vowel /ɑ/ and 9.3 - 11.0 for utterances containing the vowel /i/, depending upon the speaker. The constant compensates for the difference in rates of movement of the tongue tip and the point a few millimeters behind the tip. The point behind the tip moves more slowly, particularly for the vowel /i/. The difference in constant values for /ɑ/ and /i/ could be attributed to the fact that the majority of the tongue blade does not need to move down to produce the vowel /i/, so the point on the tongue where the transducer is located probably makes very little movement when compared to the tongue tip. In contrast, both the tongue tip and the transducer location on the tongue move down considerably during the production of /ɑ/.

Table 4.2 summarizes, for each utterance and speaker, the linear rate of increase in labial and alveolar constriction cross-sectional area following the stop release. An observation from the table is that a faster rate of area increase is associated with the labial than alveolar stop. This observation holds both within and across speakers. The rate of area increase for the labial stop is in the range 35.2 - 53.3 cm^2/sec. The

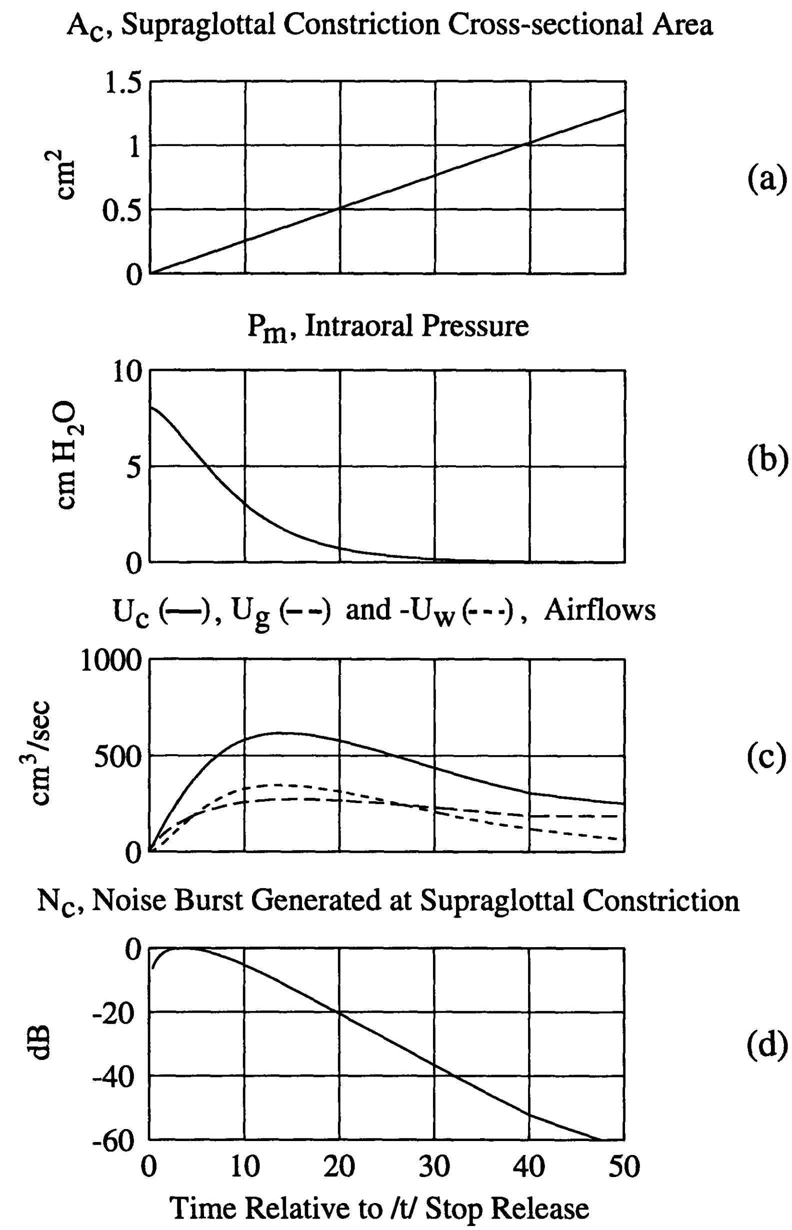

Figure 4-7: Low-frequency circuit and frication noise burst model inputs and outputs for the unaspirated stop consonant /t/ upon release of closure in the utterance stot, based on data derived from Subject 3: (a) Lip-opening constriction cross-sectional area, A_c; (b) Pressure within the mouth, P_m; (c) Airflow through the lip-opening constriction, U_c (solid line), airflow through the glottis, U_g (dashed line), and airflow generated by the inward displacement of the vocal-tract walls, -U_w (dotted line) (the negative sign indicates the direction of displacement of U_w is inward); (d) Frication noise burst, N_c. Time zero is the instant of stop release.

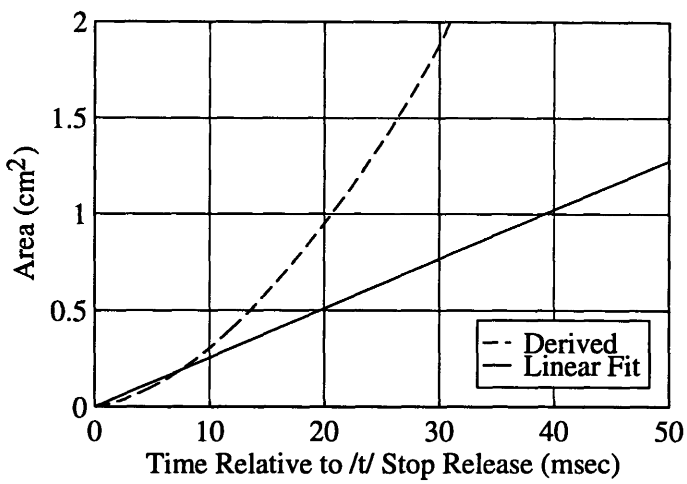

Figure 4-8: Alveolar constriction cross-sectional area for the utterance stot, based on data derived from Subject 3. The dashed line is the area derived from the articulation data, models, and acoustic data, as discussed in the text. The solid line is the best linear fit to the first 10 msec of the derived area. The linear rate of area increase is 25.5 cm^2/sec.

Utterance	Constriction Cross-sectional Area Rate of Increase (cm^2/sec)		
	Subject 1	Subject 2	Subject 3
spot	46.6	38.0	39.0
stot	24.7	-	25.5
speet	53.3	40.4	35.2
steet	29.9	-	19.6

Table 4.2: Labial and alveolar constriction cross-sectional area linear rates of increase following the unaspirated stop-consonant release.

average rate of area increase for the labial stop is 42 cm^2/sec (averaged across all utterances and speakers). The alveolar stop has a rate in the range 19 - 30 cm^2/sec, with an average rate, across all utterances and speakers, of 25 cm^2/sec. In comparison, Fujimura (1961b) measured a rate of 100 cm^2/sec for the labial stop, more than double the maximum rate determined by the present study. Stevens (Acoustic Phonetics, in preparation) derived rates of 50 - 100 cm^2/sec for an apical stop consonant, based on the work of Kent and Moll (1972), Kuehn and Moll (1976), and from acoustic data. The rates derived by Stevens are on the order of two to three times larger than the rates found in the present study.

The trends observed from Table 4.2 are consistent with the synthesis study performed by Williams (1995) (cf. Chapter 1, Section 1.2.2). Williams found that the listeners indicated a strong preference for faster rates of increase (25, 50 and 100 cm^2/sec) for the labial stop, over rates of 10 and 15 cm^2/sec. For the alveolar stop, the preference was strongly for the 25 cm^2/sec rate. Williams concluded that it is possible that a single constriction cross-sectional area rate of increase following the stop release (about 30 - 40 cm^2/sec) could be used to synthesize all voiced stops. The findings in this thesis provide some experimental support for a similar conclusion for the voiceless stops.

4.5 First Formant Frequency Transition

The resonances of the portion of the vocal tract located in front of the constriction are the principal vocal-tract resonances excited by the frication noise burst following the voiceless stop consonant release. The first formant frequency, $F1$, of the vocal tract is not excited, and the $F1$ transition is not evident in the acoustic waveform during the first 10 - 15 msec following the release. The linearized constriction cross-sectional area change with time following the stop release, as determined in Section 4.4, will be utilized in a model similar to the models presented in Chapter 2, Section 2.2, to estimate the $F1$ transition during the frication noise burst time period. The estimated $F1$ transition will supplement the acoustic data for the first several milliseconds after the release. Estimates of the $F1$ transition will be calculated only for utterances containing the vowel /ɑ/, since there is very little observable $F1$ transition for /i/.

4.5.1 Procedure

The software program KLSPEC was used to measure the $F1$ transition from the acoustic data. A 6.4 msec Hamming window was chosen to give a window duration less than the length of one glottal pulse. (For the subjects involved in this study, all male speakers, a glottal pulse typically ranged from 8.5 - 10 msec in length.) This choice of window duration captured the effects of an individual glottal pulse on the resonances of the vocal tract. The window was centered over the first glottal pulse, or pitch period, in the acoustic waveform following the stop release, typically 15 - 30 msec after the release for the speakers in this study. Next, the discrete Fourier transform was calculated for the windowed, first-differenced segment of the waveform. The peak corresponding to $F1$ was located in the spectrum using the judgment of the author. The frequency of the peak and its associated time, relative to the stop release, were recorded for that particular window location. The window was moved from pitch period to pitch period through the speech time waveform, centering the window over the pitch period each time. At each window location, the frequency of the peak and the associated time, relative to the stop release, were recorded. In this manner, the $F1$ values after the stop release were measured from the acoustic data. In addition, a value of $F1$ for the unaspirated stops at the instant of release is available from the literature (Refer to Chapter 2, Section 2.2). $F1$ is initially assigned the value 210 Hz, based on experiment and theory. This value represents the effects on the vocal-tract impedance of the nonrigid walls and partially open glottis.

To estimate the $F1$ transition during the production of the frication noise burst, a time interval when $F1$ is not evident in the acoustic data, the linear rate of constriction cross-sectional area increase following release is incorporated into models of the vocal-tract shape for /p/ and /t/. The models are similar to the vocal-tract tube models discussed in Chapter 2, Section 2.2.

The vocal-tract model for the unaspirated /p/ preceding the vowel /ɑ/ is shown in Figure 4-9. The labial constriction is represented by a short tube at the right end of the model. The vertical arrows above and below that tube indicate that the constriction cross-sectional area increases with time following the release. The linear rate of area increase determined in Section 4.4 for /p/ is used as the cross-sectional area of that tube. The remainder of the vocal tract, to the left of the constriction, is fixed in the configuration required for the upcoming vowel /ɑ/. The model assumes that this portion of the vocal tract has anticipated the upcoming vowel by the time of the stop release. The vocal-tract model for the unaspirated /t/ preceding the vowel /ɑ/ is shown in Figure 4-10. The alveolar constriction is represented by two adjacent sections in the model, having lengths L_{m2} and L_c in the figure. The vertical

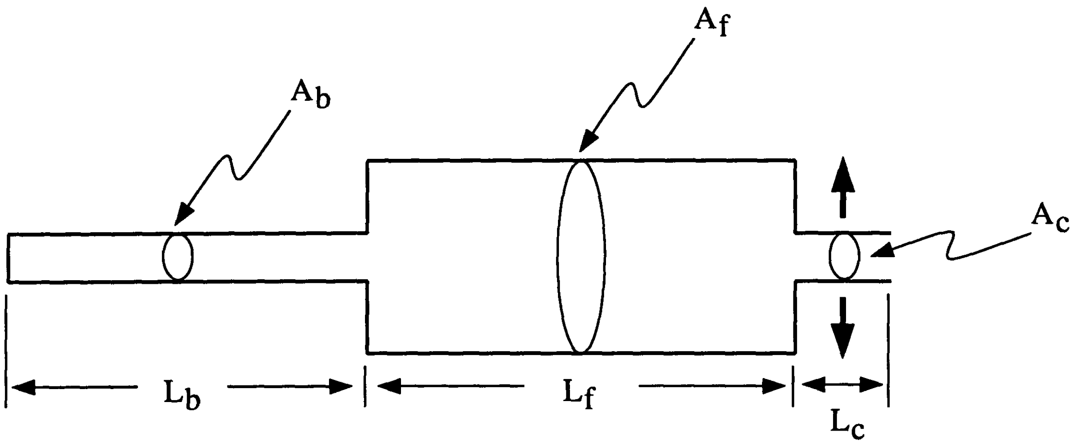

Figure 4-9: Vocal-tract tube model for production of the unaspirated stop consonant /p/ following the release. L_b is the length of the back cavity, A_b is the cross-sectional area of the back cavity, L_f is the length of the front cavity, and A_f is the cross-sectional area of the front cavity. These two cavities are fixed in the configuration required for the upcoming vowel /ɑ/. L_c is the length of the constriction, about 5 mm for /p/, and A_c is the constriction cross-sectional area, increasing at the linear rate obtained in Section 4.4.

arrows above and below the alveolar constriction in the model indicate that the cross-sectional area increases with time following the release. The linear rate of area increase determined in Section 4.4 for /t/ is used as the rate of increase in cross-sectional area of the short tube comprising the right portion of the alveolar constriction (the tube with length L_c). For the portion of the constriction which gradually narrows from left to right (the section of tube having length L_{m2}), the rate of increase is adjusted on a sliding scale from 0 at the left edge of the section to the full linear rate at the right edge. The remainder of the vocal tract, to the left and right of the constriction (the portions of the vocal tract with lengths L_b, L_{m1} and L_f), is fixed in the configuration required for the upcoming vowel /ɑ/. The model assumes that the vocal tract, with the exception of the constriction, anticipates the following vowel by the time of the stop release[1].

Maeda's program (described in Chapter 2, Section 2.2), when given the cross-sectional area of the entire vocal tract at a particular instant in time, solves the

[1]This alveolar model has only a partially-fronted tongue body position, but it is believed to be adequate for estimating $F1$.

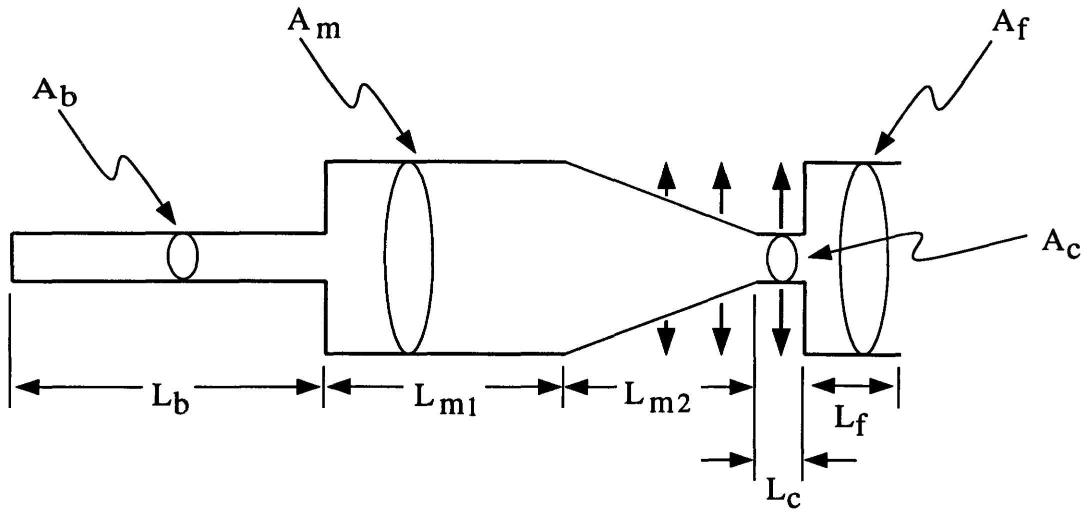

Figure 4-10: Vocal-tract tube model for production of the unaspirated stop consonant /t/ following the release. L_b is the length of the back cavity, A_b is the cross-sectional area of the back cavity, L_{m1} is the length of the first middle cavity, A_m is the cross-sectional area of the first middle cavity, L_f is the length of the front cavity, and A_f is the cross-sectional area of the front cavity. These three cavities are fixed in the configuration required for the upcoming vowel /ɑ/. L_c is the length of the most narrow portion of the constriction, about 5 mm for /t/, and A_c is the cross-sectional area of that portion of the constriction, increasing at the linear rate obtained in Section 4.4. The portion of the model with length L_{m2} represents the remainder of the alveolar constriction and has a gradually increasing rate of cross-sectional area increase, from 0 at the left edge to the full linear rate from Section 4.4 at the right edge.

wave equation to yield the corresponding value of $F1$. By gradually increasing the constriction cross-sectional area between program runs, $F1$ as a function of time following the stop release can be determined. Except for the constriction region, the lengths and cross-sectional areas of the vocal tract are based on the vocal-tract configuration required for the production of the vowel /ɑ/ (Fant (1960)), physiological constraints, and an attempt to obtain agreement between the $F1$ transition predicted by Maeda's program and the $F1$ transition measured from the acoustic data. The length of the constriction was 0.5 cm and the constriction cross-sectional area was given by the linear rate of area increase determined in Section 4.4. In the models of Figures 4-9 and 4-10 the glottis is on the left, and is modeled as closed. Maeda's program does not have the capability to incorporate an opening in the glottis, such as the opening which exists during the production of an unaspirated stop. Consequently, only the effect of the nonrigid walls is seen in the value of $F1$ at the time of the release.

A simpler approach to estimating the $F1$ transition immediately following the release would be to fit an exponentially-rising curve to the measured $F1$ values obtained from the acoustic data (including the $F1$ value of 210 Hz at the time of release). The exponential curve fit to the data was of the form $C_1 + C_2(1 - e^{-t/\tau})$, where $C_1 = 210$ Hz. The constants C_2 and τ were determined by minimizing the mean square error between the curve and the $F1$ values over the first 50 msec following the release for all repetitions of a given utterance for a particular speaker. Comparison of the resultant curve to the $F1$ transition predicted by Maeda's program will determine if an exponential curve provides a reasonable approximation to the $F1$ transition.

4.5.2 Results and Discussion

The individual first formant frequency ($F1$) values, measured from the acoustic data, appear in Figure 4-11 for the time period following the stop-consonant release in the utterance spot, spoken by Subject 3. Measured values of $F1$ are shown for all of the repetitions of the utterance. The $F1$ transition generated by Maeda's program, which incorporated the linear rate of area increase of 39.0 cm^2/sec determined in Section 4.4, is also displayed in the figure. The first few measured values of $F1$ do not appear until approximately 10 - 15 msec following the release in this utterance. This time period is supplemented with the $F1$ transition generated by Maeda's program. The parameters used in the program for the vocal-tract tube segment lengths and cross-sectional areas are $L_b = 7.5$ cm, $A_b = 1.0$ cm^2, $L_f = 9.0$ cm, $A_f = 5.5$ cm^2, and $L_c = 0.5$ cm (see Figure 4-9). The total vocal-tract length used by the model, therefore, is 17.0 cm, a typical vocal-tract length for male speakers. Parameters for the exponential curve fit were found to be $C_2 = 413$ and $\tau = 10.3$ msec.

Figure 4-12 shows the $F1$ transitions, measured and theoretical, for the utterance stot, spoken by Subject 3. The Maeda program parameter values for the alveolar vocal-tract tube model are $L_b = 6.5$ cm, $A_b = 0.7$ cm^2, $L_{m1} = 4.0$ cm, $A_m = 5.6$ cm^2, $L_{m2} = 4.0$ cm, $L_c = 0.5$ cm, $L_f = 2.0$ cm, and $A_f = 5.6$ cm^2 (see Figure 4-10). Parameters for the exponential curve fit were found to be $C_2 = 472$ and $\tau = 18.8$

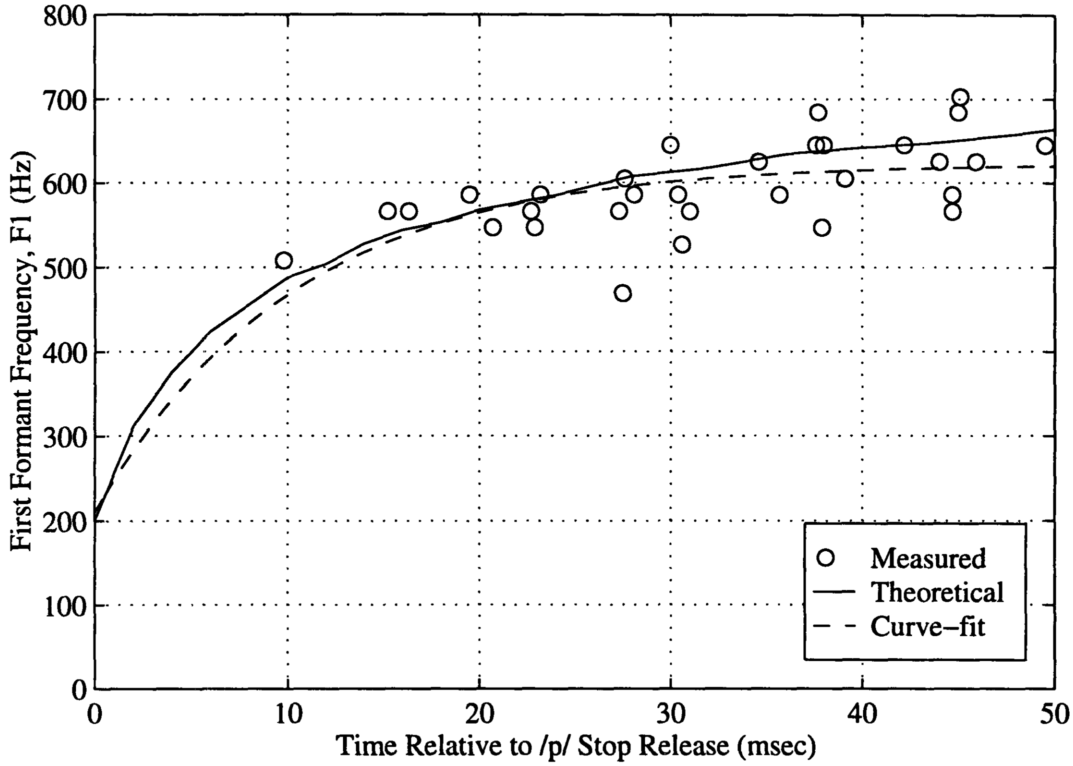

Figure 4-11: First formant frequency ($F1$) transition following the release of the unaspirated /p/ in spot, spoken by Subject 3. Dots represent values of $F1$ measured from the acoustic data for all repetitions. The solid line is the theoretical estimate of the $F1$ transition, generated by Maeda's program which incorporated linear area changes estimated from articulation and acoustic data. The dashed line is the exponential curve fit to the measured acoustic data. See text for theoretical and curve-fit parameter values.

msec.

Figure 4-13 shows the $F1$ transitions, measured and theoretical, for the utterance spot, spoken by Subject 1. (Refer to the example in Chapter 2, Section 2.2.) The Maeda program parameter values for the labial vocal-tract tube model are $L_b = 7.0$ cm, $A_b = 1.5$ cm^2, $L_f = 9.5$ cm, $A_f = 5.5$ cm^2, and $L_c = 0.5$ cm (refer to Figure 4-9). The total vocal-tract length is 17.0 cm for this speaker as well. In order to better match the $F1$ values in the vowel, these parameter values varied slightly from those of Subject 3 for the same utterance. Parameters for the exponential curve fit were found to be $C_2 = 440$ and $\tau = 16.2$ msec.

Comparison of the $F1$ transitions following the unaspirated stop release in spot

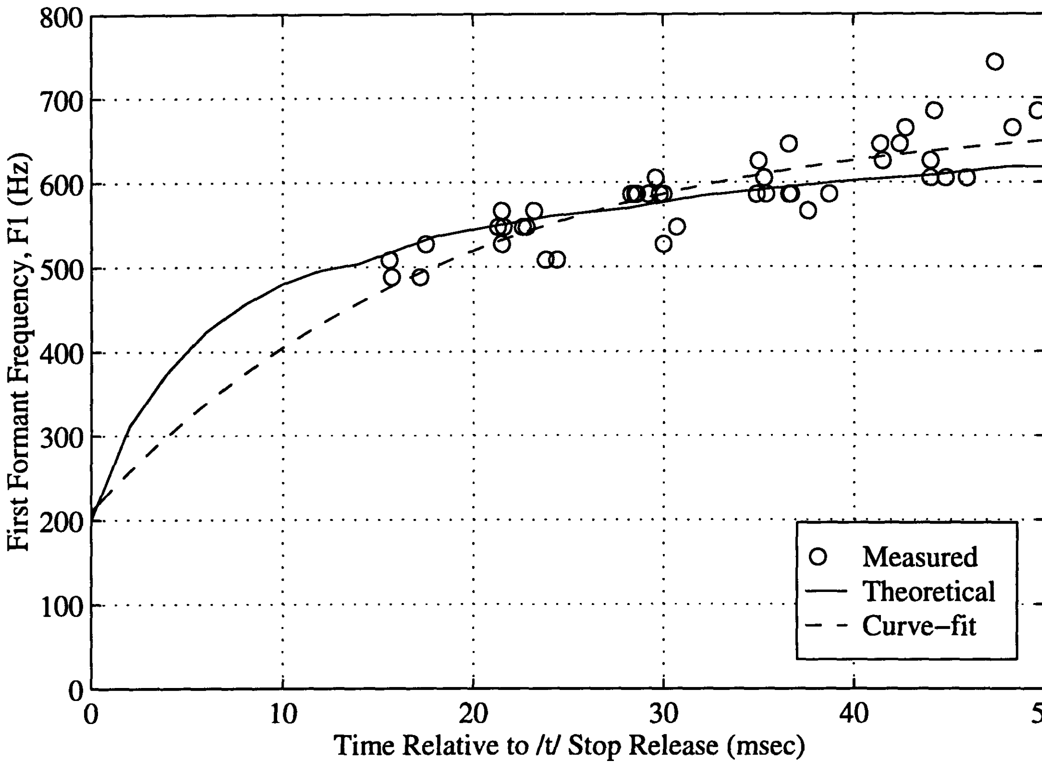

Figure 4-12: First formant frequency ($F1$) transition following the release of the unaspirated /t/ in stot, spoken by Subject 3. Dots represent values of $F1$ measured from the acoustic data for all repetitions. The solid line is the theoretical estimate of the $F1$ transition, generated by Maeda's program which incorporated linear area changes estimated from articulation and acoustic data. The dashed line is the exponential curve fit to the measured acoustic data. See text for theoretical and curve-fit parameter values.

and stot shows that the rates of rise in $F1$ immediately following the stop release are very similar for Subject 3. For Subject 1, the rate of rise is slightly slower following the unaspirated /t/ release in stot (not shown) than for the unaspirated /p/ release in spot. The rates of rise are virtually identical across all speakers for the utterance spot, but the rate of rise is slower for stot spoken by Subject 1 than spoken by Subject 3.

The $F1$ transition calculated by Maeda's program, utilizing the labial and alveolar constriction cross-sectional area linear rates of increase following the release, fits the measured $F1$ values reasonably well for each utterance and speaker. The ability of the calculated $F1$ transition to approximately match the $F1$ transition measured

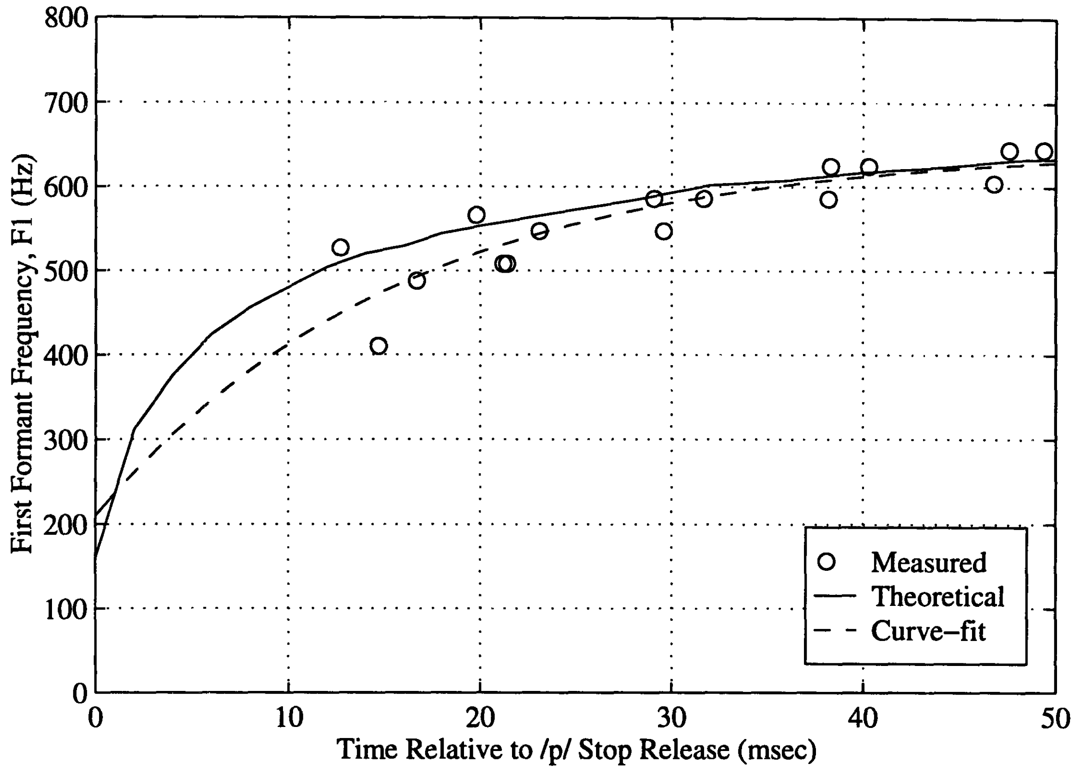

Figure 4-13: First formant frequency ($F1$) transition following the release of the unaspirated /p/ in <u>spot</u>, spoken by Subject 1. Dots represent values of $F1$ measured from the acoustic data for all repetitions. The solid line is the theoretical estimate of the $F1$ transition, generated by Maeda's program which incorporated linear area changes estimated from articulation and acoustic data. The dashed line is the exponential curve fit to the measured acoustic data. See text for theoretical and curve-fit parameter values.

is a reasonable approach to determining the rate of constriction cross-sectional area change with time after the release. The $F1$ transition determined by Maeda's program has successfully supplemented the $F1$ transition measured from the acoustic data for the frication noise burst time period.

An exponential curve fit to the acoustic data is shown in each of Figures 4-11, 4-12, and 4-13. The advantage of using such a method to approximate the $F1$ transition is that it relies solely on knowledge of the acoustic data. When compared to the $F1$ transition determined by Maeda's program, the curve fit may be considered somewhat accurate for the initial portion of the $F1$ transition following the /p/ release in <u>spot</u>, as spoken by Subjects 2 (not shown) and 3. It is evident, however, that the initial rate of rise in the exponential curve is not fast enough to accurately represent the $F1$ transition during the initial 20 - 25 msec following the release in spot, spoken by Subject 1. Similarly, the exponential curve does not provide a good approximation to the $F1$ transition for the initial 30 - 35 msec following the release in stot spoken by Subject 1 (not shown), and for the initial 20 msec after the release in stot spoken by Subject 3. As might be expected, when there are fewer measured $F1$ values during the first 50 msec following the release, as in the case of a longer VOT, the exponential curve fit is a poorer approximation. The fact that the curve fit does not accurately represent the $F1$ transition following release for all utterances and speakers is evidence that a technique such as the one described in this chapter, using articulation data, acoustic data, modeling and Maeda's program to determine the $F1$ transition, provides better estimates of the transition.

4.6 Summary

Section 4.1 provides a description of the speakers' backgrounds and the selected corpus. Section 4.2 discusses the acoustic recording method utilized in the study.

In Section 4.3, the frication noise burst was measured in the acoustic data using the software program KLSPEC. The duration of the noise burst, the time interval lying between the 10 dB down points from its peak amplitude, was recorded for each utterance and speaker. The duration was shorter for utterances containing an unaspirated /p/ (spot, speet) than utterances containing an unaspirated /t/ (stot, steet).

In Section 4.4, an initial approximation to the constriction cross-sectional area change with time following the release was derived from the articulation data. An ellipse was used to model the labial constriction and the segment of a circle to model the alveolar constriction. A linear rate of area increase was estimated from the first 10 msec of the modeled cross-sectional area. The linear rate was used as an input parameter to the low-frequency circuit model of Chapter 2, Section 2.1. The airflow through the constriction, an output of the circuit model, was used as an input to a model of the frication noise burst generated at the constriction following the stop release. The 10 dB down duration of the modeled noise burst was compared to the noise burst duration measured from the acoustic data in Section 4.3. The modeled cross-sectional area, and associated linear rate of area increase, were adjusted until the linear rate produced a modeled noise burst equal in duration to the measured burst. The resulting linear rates of area increase revealed a faster rate for utterances containing the unaspirated /p/ than the unaspirated /t/.

In Section 4.5, first formant frequency values were measured from the acoustic data. An estimate of the transition was also determined using the linear rate of area increase derived in Section 4.4 as a parameter in a model of the vocal-tract configuration. Given the cross-sectional area over the length of the vocal tract, Maeda's program was utilized to solve the wave equation for the $F1$ transition. The $F1$ transition generated by Maeda's program supplemented the acoustic data during the frication noise burst time period when $F1$ was not excited by a vocal-tract source. During the time interval when acoustic data was available following the release, the $F1$ transition generated by Maeda's program represented the measured $F1$ values well, indicating that the approach used in Section 4.4 to estimate the linear rate of constriction cross-sectional area increase following the release was a reasonable approach. It was shown that an exponential curve does not consistently provide a good estimate of the $F1$ transition.

In summary, the articulation data, coupled with the use of models, was used to supplement the acoustic data, providing information about the articulator movements and the resultant acoustics immediately following the stop-consonant release. This study is one of very few which correlates articulatory and acoustic events in order to determine more about stop-consonant production.

Chapter 5

Conclusion

5.1 Summary of Results

This thesis investigated the production of the labial and alveolar voiceless stop consonants through examination of experimentally-recorded articulation and acoustic data. Articulator movements, as measured by small transducers placed on the individual articulators, were recorded by an electro-magnetic midsagittal articulometer (EMMA) system. The acoustic signal was recorded simultaneously. Several phonetic contexts for the stop consonants were examined. For comparison purposes, the remaining stop and nasal consonants were also recorded.

In the first part of the study, a data processing procedure which incorporated the use of events observed in the acoustic signal, such as the stop release, was developed to average the articulator displacement waveforms for a given utterance and speaker. To account for the presence of slight variations in speaking rate between repetitions, the procedure employed linear time warping. Analysis of the average articulator displacement waveforms and the average maximum downward velocities following the stop release yielded a few observations: (1) The average rate of maximum downward jaw movement following the time of the stop release is greater by 1.4 cm^2/sec, on average, for utterances in which the stop is preceded by /s/. The jaw is constrained to maintain a high position until the end of /s/ production. The jaw then moves rapidly downward in an attempt to compensate for the high position and produce the following vowel in a timely manner; (2) When a stop or nasal consonant is followed by the vowel /ɑ/, the jaw moves downward more rapidly and has a greater relative displacement than when followed by /i/. This observation provides evidence of a correlation between larger articulator displacements and faster rates of movement; (3) Examining lower lip movement during the production of /p/ in pot, one of the speakers in the study, Subject 2, has the strategy of making a larger movement, corresponding to a faster maximum velocity following the release, than the other two speakers. Subject 2 also takes more time than the other speakers to produce the stop, possibly indicating incomplete compensation for the larger movement; (4) A bandwidth of approximately 3 - 5 Hz exists for cyclic articulatory movements produced in speech, such as the movement of the lower jaw from the steady-state portion of the vowel /e/ in say to the steady-state portion of the vowel /ɑ/ in pot, spoken in the phrase, "Say pot again."; and (5) The duration of time a primary articulator takes to produce a stop (including the onset phase, closure interval and offset phase) is about 200 - 300 msec, equivalent to the length of one period of the cyclic movement.

In the second part of the study, linear estimates of the rates of labial and alveolar constriction cross-sectional area increase following the release were determined from the articulation data, with the aid of models and the acoustic data. First, the frication noise burst following the stop release was measured from the acoustic data. The duration of the noise burst, defined to be the time interval between the 10 dB down points from the peak noise amplitude, was recorded for each utterance and speaker. The duration was observed to be shorter for utterances containing an unaspirated /p/, in the range 7 - 10 msec, than utterances containing an unaspirated /t/, 11 - 16 msec. Second, an initial approximation to the constriction cross-sectional area change with time following the release was derived from the articulation data. An ellipse was used to model the labial constriction and the segment of a circle to model the alveolar constriction. A linear rate of area increase was estimated from the first 10 msec of the modeled cross-sectional area. Third, the linear rate was used as an input parameter to a low-frequency circuit model of the average pressures and airflows in the vocal tract during stop production. Fourth, the airflow through the constriction, an output of the circuit model, was used as an input to a model of the frication noise burst generated at the constriction following the stop release. Fifth, the 10 dB down duration of the modeled noise burst was compared to the noise burst duration measured from the acoustic data. Lastly, the modeled cross-sectional area, and the associated linear rate of area increase, were adjusted iteratively until the linear rate produced a modeled noise burst equal in duration to the measured burst. The resulting linear rates of area increase were in the range of approximately 35 - 50 cm^2/sec for the voiceless labial stop, with an average of 42 cm^2/sec, and a range of approximately 20 - 30 cm^2/sec for the voiceless alveolar stop, with an average of 25 cm^2/sec. The results indicated a faster rate of opening was utilized for utterances containing the unaspirated /p/ than the unaspirated /t/.

An estimate of the first formant frequency ($F1$) transition during production of the frication noise burst was determined by incorporating the linear rate of constriction cross-sectional area increase into a high-frequency vocal tract model. Given the cross-sectional area over the length of the vocal tract, a program (written by S. Maeda) solved the wave equation to yield an estimate of the $F1$ transition following the stop release. First formant frequency values were also measured from the acoustic data. The estimated $F1$ transition supplemented the acoustic data during the frication noise burst time period when $F1$ was not excited by a vocal-tract source. During the time interval when acoustic data was available following the release, the $F1$ transition generated by Maeda's program represented the measured $F1$ values well, indicating that the approach used to estimate the linear rate of constriction cross-sectional area increase following the release was reasonable. It was also shown that an exponential curve fit to the acoustic data does not consistently provide a good estimate of the $F1$ transition.

5.2 Directions for Future Research

In this study, the estimate of the constriction cross-sectional area change with time following the release was not able to be derived solely from the articulator movements recorded by the transducers. The use of acoustic data was required to develop a reasonable estimate of the area rate. Even with the use of the acoustic data, the area rate still had to be linearized, which is probably not a good representation of the changing shape of the constriction within the first 1 - 2 msec following the release. Experimental devices which are capable of directly measuring changes in the constriction cross-sectional area at a high frame rate need to be developed.

A natural extension to the study would be the inclusion of the velar voiceless stop. Consideration would need to be given to the shape chosen to represent the constriction cross-sectional area and to the position of the tongue body transducer compared to the constriction location, since the two positions could differ substantially at times.

The present study is one of a select few utilizing both articulation and acoustic data to investigate stop consonant-production. The acoustic signal is a source of information about speech production which is often overlooked. More studies need to investigate and address the relationship between articulation and acoustic data.

5.3 Implications for Speech Recognition, Speech Synthesis, and Analysis of Disordered Speech Production

A speech recognition system will be able to recognize a larger database of utterances and speakers through improved understanding of the influences of phonetic context and interspeaker variability on stop production.

Speech synthesizers which use information derived from articulator movements will benefit by experimentally-based knowledge of the range of constriction cross-sectional area rates of increase following the stop release.

The results of this study could also be used as a baseline for comparison to disordered speech production. A patient's speech is examined by a speech pathologist to aid in the diagnosis and remediation of a speech production disorder. The present study relates the acoustic data to the corresponding articulatory movements. Consequently, if a patient's speech was compared to the acoustic data in the study, it may be possible to determine the ways in which the articulators are moving incorrectly.

Bibliography

[1] S. M. Barlow, K. J. Cole, and J. H. Abbs. A new head-mounted lip-jaw movement transduction system for the study of motor speech disorders. *J. Speech and Hear. Res.*, 26:283–288, 1983.

[2] M. E. Beckman and J. Edwards. Articulatory evidence for differentiating stress categories. In P. A. Keating, editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 7–33. Cambridge: Cambridge University Press, 1993.

[3] L. L. Beranek. *Acoustics.* Wiley, New York, 1954.

[4] G. J. Borden and T. Gay. Temporal aspects of articulatory movements for /s/-stop clusters. *Phonetica*, 36:21–31, 1979.

[5] M. F. Dorman and L. J. Raphael. Distribution of acoustic cues for stop consonant place of articulation in VCV syllables. *J. Acoust. Soc. Am.*, 67(4):1333–1335, 1980.

[6] J. Edwards. Contextual effects on lingual-mandibular coordination. *J. Acoust. Soc. Am.*, 78(6):1944–1948, 1985.

[7] O. Engstrand. Acoustic constraints or invariant input representation? An experimental study of selected articulatory movements and targets. Reports from Uppsala University Department of Linguistics 7, Uppsala University, Uppsala, Sweden, 1980.

[8] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

[9] G. Fant. Vocal tract wall effects, losses, and resonance bandwidths. Technical report, Speech Transmission Laboratory Quarterly Progress and Status Report 2-3, Royal Institute of Technology, Stockholm, Sweden, 1972.

[10] G. Fant, L. Nord, and P. Branderud. A note on the vocal tract wall impedance. Technical report, Speech Transmission Laboratory Quarterly Progress and Status Report 4, Royal Institute of Technology, Stockholm, Sweden, 1976.

[11] O. Fujimura. Effects of vowel context on the articulation of stop consonants. Acoustical Society of America Conference, May 1961a.

[12] O. Fujimura. Bilabial stop and nasal consonants: A motion picture study and its acoustical implications. *J. Speech and Hear. Res.*, 4(3):233–247, 1961b.

[13] O. Fujimura and J. Lindqvist. Sweep-tone measurements of vocal-tract characteristics. *J. Acoust. Soc. Am.*, 49:541–558, 1971.

[14] T. Gay. Articulatory movements in VCV sequences. *J. Acoust. Soc. Am.*, 62(1):183–193, 1977.

[15] J. Glass. Electrical Engineering and Computer Science Departmental Area Exam. Massachusetts Institute of Technology, Cambridge, MA, 1986.

[16] V. L. Gracco. Characteristics of speech as a motor control system. In G. Hammond, editor, *Cerebral control of speech and limb movements*, pages 3 – 28. North Holland: Elsevier, 1990.

[17] V. L. Gracco. Sensorimotor mechanisms in speech motor control. In H. Peters, W. Hultsijn, and C. W. Starkweather, editors, *Speech motor control and stuttering*, pages 53 – 78. North Holland: Elsevier, 1991.

[18] V. L. Gracco. Some organizational characteristics of speech movement control. *J. Speech and Hear. Res.*, 37:4–27, 1994.

[19] V. L. Gracco and A. Löfqvist. Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. Technical Report SR - 115/116, Haskins

Laboratories Status Report on Speech Research, July - December 1993. To appear in Journal of Neuroscience, 1994.

[20] W.L. Henke. *MITSYN Languages, Language Reference Manual.* WLH, 133 Bright Rd., Belmont, MA 02178, 1989.

[21] C. Henton, P. Ladefoged, and I. Maddieson. Stops in the world's languages. *Phonetica*, 49:65–101, 1992.

[22] T. Ichikawa, J. Komoda, M. Horiuchi, and N. Matsumoto. Characteristics of articulatory coordination during japanese VCV sequences: Observations using optical and marker-tracking systems. *J. Acoust. Soc. Jpn. (E)*, 14:1–9, 1993.

[23] K. Ishizaka, J. C. French, and J. L. Flanagan. Direct determination of vocal tract wall impedance. In *IEEE Trans. on Acoust., Speech and Signal Processing,* ASSP-23, pages 370–373, 1975.

[24] R. D. Kent and K. L. Moll. Cinefluorographic analyses of selected lingual consonants. *J. Speech and Hear. Res.*, 15:453–473, 1972.

[25] D. P. Kuehn and K. N. Moll. A cineradiographic study of VC and CV articulatory velocities. *J. Phonetics*, 4:303–320, 1976.

[26] A. M. Liberman, P. C. Delattre, and F. S. Cooper. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167, 1958.

[27] R. N. Linville. *Temporal Aspects of Articulation: Some Implications for Speech Motor Control of Stereotyped Productions.* PhD thesis, University of Iowa, Iowa City, IA, 1982.

[28] L. Lisker and A. S. Abramson. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20:384–422, 1964.

[29] L. Lisker and A. S. Abramson. The voicing dimension: Some experiments in comparative phonetics. In *Proc. 6th Int. Cong. Phonetic Sciences, Prague 1967*,

pages 563–567, Prague, 1970. Academia Publ. House of Czechoslovak Acad. of Science.

[30] M. Macchi. Labial articulation patterns associated with segmental features and syllable structure in English. *Phonetica*, 45:109–121, 1988.

[31] P. M. Morse. *Vibration and sound.* McGraw-Hill, New York, 1948.

[32] E. M. Müller and J. H. Abbs. Strain gauge transduction of lip and jaw motion in the midsagittal plane: Refinement of a prototype system. *J. Acoust. Soc. Am.*, 65:481–486, 1979.

[33] E. M. Müller and W. S. Brown, Jr. Variations in the supraglottal air pressure waveform and their articulatory interpretations. In N. J. Lass, editor, *Speech and Language: Advances in Basic Research and Practice*, volume 4, pages 317–389. Academic Press, New York, 1980.

[34] W. L. Nelson, J. S. Perkell, and J. R. Westbury. Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations. *J. Acoust. Soc. Am.*, 75(3):945–951, 1984.

[35] Z. B. Nossair and S. A. Zahorian. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *J. Acoust. Soc. Am.*, 89(6):2978–2991, 1991.

[36] R. N. Ohde. Fundamental frequency as an acoustic correlate of stop consonant voicing. *J. Acoust. Soc. Am.*, 75(1):224–230, 1984.

[37] R. N. Ohde and K. N. Stevens. Effect of burst amplitude on the perception of stop consonant place of articulation. *J. Acoust. Soc. Am.*, 74(3):706–714, 1983.

[38] D. J. Ostry and K. G. Munhall. Control of rate and duration of speech movements. *J. Acoust. Soc. Am.*, 77(2):640–647, 1985.

[39] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network

trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, 92(2, Pt. 1):688–700, 1992.

[40] L. Pastel. Turbulent noise sources in vocal tract models. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[41] J. S. Perkell. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* MIT Press, Cambridge, MA, 1969.

[42] J. S. Perkell. Articulatory processes. In W. J. Hardcastle and J. Laver, editors, *A Handbook of Phonetic Science.* 1994.

[43] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J. Acoust. Soc. Am.*, 92(6):3078–3096, 1992.

[44] M. Rothenberg. The breath stream dynamics of simple-released-plosive production. Bibliotheca Phonetica No. 6, S. Karger, Basel, 1968.

[45] E. L. Saltzman. Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer and C. Fromm, editors, *Generation and modulation of action patterns*, pages 129–144. Springer-Verlag, Berlin, 1986.

[46] C. Shadle. The acoustics of fricative consonants. RLE Technical Report 506, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[47] B. L. Smith and T. E. Gartenberg. Initial observations concerning developmental characteristics of labio-mandibular kinematics. *J. Acoust. Soc. Am.*, 75(5):1599–1605, 1984.

[48] B. L. Smith and A. McLean-Muse. Articulatory movement characteristics of labial consonant productions by children and adults. *J. Acoust. Soc. Am.*, 80(5):1321–1328, 1986.

[49] K. N. Stevens. Acoustic Phonetics. Book, in preparation.

[50] K. N. Stevens. Airflow and turbulence noise for fricative and stop consonants. *J. Acoust. Soc. Am.*, 50:1180–1192, 1971.

[51] K. N. Stevens. Models for the production and acoustics of stop consonants. *Speech Comm.*, 13:367–375, 1993.

[52] K. N. Stevens and A. S. House. Studies of formant transitions using a vocal tract analog. *J. Acoust. Soc. Am.*, 28:578–585, 1956.

[53] K. N. Stevens and D. H. Klatt. Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am.*, 55(3):653–659, 1974.

[54] H. M. Sussman, P. F. MacNeilage, and R. J. Hanson. Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *J. Speech and Hear. Res.*, 16:397–420, 1973.

[55] J. R. Westbury. Aspects of the temporal control of voicing in consonant clusters in English. Texas Linguistic Forum 14, Department of Linguistics, University of Texas, 1979.

[56] D. R. Williams. Perception of oral release rate for initial voiced stops. Stockholm, Sweden, August 1995. XIIIth International Congress of Phonetic Sciences. Submitted April 15, 1995. Scheduled to appear in Conference Proceedings.

[57] G. R. Wodicka, A. M. Lam, and V. Bhargara. Acoustic impedance of the maternal abdomen. *J. Acoust. Soc. Am.*, 94:13–18, 1993.

[58] V. W. Zue. Acoustic characteristics of stop consonants: A controlled study. Technical report, Indiana University Linguistics Club, Indiana University, 310 Lindley Hall, Bloomington, Indiana 47405, 1976 (first reproduced in 1980).