THE CONTRIBUTION OF SPEECH SYNTHESIS TO PHONETICS: DENNIS KLATT'S LEGACY

Kenneth N. Stevens

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

ABSTRACT

Some of Dennis Klatt's contributions to the science and application of speech synthesis are described, and the effects of these contributions on the study of phonetics are discussed. The synthesizer developed by Klatt, with an extended set of control parameters, can be manipulated to simulate different female and male voices and can produce a variety of classes of speech sounds in context, based on principles of human speech sound generation. The problem of controlling the multiple parameters of the synthesizer is considered, in view of the constraints imposed on the parameters by the articulatory and aerodynamic processes in speech.

1. INTRODUCTION

One of Dennis Klatt's contributions to the field of phonetics was to advance the science and application of speech synthesis. He approached the problem of speech synthesis in a systematic way, incorporating and contributing to what is known about the speech production process and collecting empirical data for situations where theoretical models were inadequate. I shall try to summarize the major contributions he made to this field, the relevance of these contributions to the study of phonetics, and some new directions in speech synthesis and indeed in phonetic theory that have been made possible because of the groundwork established by Dennis Klatt.

Some of his innovations in speech synthesis are concerned with the organization of the synthesizer itself, and others involve his development of rules for controlling the synthesizer. We shall examine first the arrangement of components in the most recent version of the Klatt synthesizer, called KLSYN88.

A block diagram of the synthesizer is shown in Fig. 1. There are four main components of this synthesizer: (1) a source simulating the glottal output, (2) a source of frication noise, (3) a transfer function for the glottal source, and (4) a transfer function for the frication source. The arrows pointing to the upper and lower sides of the boxes indicate that certain parameters of the sources and filters can be controlled by the user. Dennis Klatt has contributed significantly to several of these components and their interconnections.

2. LINKING FORMANTS FOR GLOTTAL AND FRICATION SOURCES

One of the problems for the synthesis of speech with a formant or terminal-analog synthesizer is that the nature of the transfer function from the source to the sound output at the mouth or nose is different when the source is at the glottis than when it is a transient or frication source located at one or more points along the length of the vocal tract. When the source is at the glottis, the transfer function is an all-pole function in the case of nonnasal vowels, and there may be additional poles and zeros for nasal vowels and for nasal and liquid consonants. In the case of a frication source, only certain of the natural frequencies of the vocal tract are excited, and it is possible to describe the transfer function as being characterized by free poles, free zeros, and pole-zero pairs [1].

Dennis Klatt observed that, for both types of sources, the poles are the same, being the natural frequencies of the vocal tract independent of the source location. For the frication source, these natural frequencies are excited with very different strengths, with primary excitation of the cavity in front of the constriction and of a possible palatal channel behind the constriction. He designed a synthesizer configuration which had two separate paths, as shown in Fig. 1 -- a cascade arrangement of poles and zeros for the glottal source and a parallel arrangement of resonators, with associated adjustable gains, for the frication source. The frequencies of the parallel resonators and of the poles in the cascade path are linked together, as the figure indicates, thereby incorporating the constraint that the natural frequencies change continuously independent of the source location within the vocal tract [7].

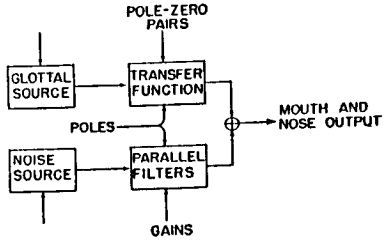

Fig. 1 Block diagram of the main components of a terminal-analog speech synthesizer such as KLSYN88 [9]. The vertical arrows on the sides of the boxes indicate arrays of control parameters.

3. SYNTHESIS OF SOUNDS WITH FRICATION NOISE

The part of the synthesizer that generates sounds with a frication source is shown in the form of a block diagram in Fig. 2. The spectrum of a sound produced with frication noise is shaped by adjusting the gains for each of the parallel formant resonators (A2F, A3F, etc.), together with a gain for a bypass path (AB) for which there is no filtering of the noise source. In order to generate noise bursts and fricative

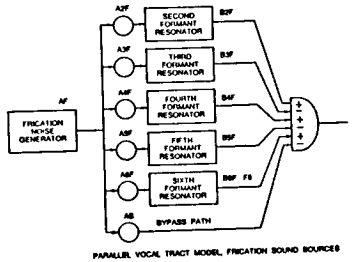

Fig. 2 Organization of the components of the KLSYN88 synthesizer for producing sounds generated with frication noise.

consonants, usually only a subset of the gains are active, since there is appreciable excitation of only some of the natural frequencies of the vocal tract. One of the projects that Klatt was working on in 1987 and 1988 was to improve the synthesis of fricative consonants. One component of this project was to synthesize frication noise with the proper spectrum. The task involved measuring the spectra of different fricative consonants in different vowel environments, produced by a male and a female speaker. In order to synthesize the fricative consonants, it was necessary to adjust the various gains for the formant filters so that the synthesized spectrum matched the original spectrum. The match could be achieved quite accurately, as Fig. 3 shows. This figure displays the spectra of two different fricative sounds produced by a male speaker, together with spectra synthesized by proper selection of the array of gains in Fig. 2. The upper two panels with the naturally produced fricatives also show smoothed spectra for a vowel adjacent to the fricative, to indicate the relative spectrum amplitudes of vowel and fricative. Comparison of the fricative spectra in the upper two panels and the synthesized spectra in the lower two panels shows reasonable agreement. In the case of the labiodental fricative, the gain AB of the bypass path was adjusted, with all other gains set to zero, whereas for the alveopalatal fricative, the gains A3F and A4F contributed the salient attributes to the spectrum. These observations are consistent with theories of fricative production [1].

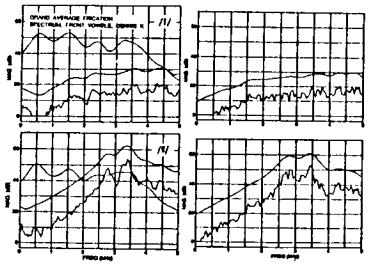

Fig. 3 The left two panels show measured spectra from syllables consisting of voiceless fricatives /f/ and /ʃ/ followed by front vowels. The fricative spectra are averages over a number of utterances, and a smoothed spectrum is displayed above each measured spectrum. The superimposed vowel spectrum is the smoothed spectrum of an adjacent vowel /ɛ/. The right two panels show attempts at matching the two left fricative spectra by setting the gains in the system in Fig. 2.

For some fricatives, particularly /s/ and /ʃ/, there are some differences in the spectrum depending on the vowel environment. In a rule system for speech synthesis in English, Klatt used somewhat different arrays of gains for fricative synthesis depending on whether the adjacent vowel was in one of three classes: front vowels, unrounded back vowels, and rounded back vowels.

4. SYNTHESIS OF NASAL CONSONANTS AND VOWELS

Synthesis of oral vowels is achieved in the usual way with an all-pole system consisting of a cascade of resonators. For the synthesis of nasal vowels and consonants, Klatt introduced a pole-zero pair into the cascade path, in addition to the poles normally involved in the generation of nonnasal vowels. The arrangement of poles and pole-zero pairs in this cascade path is shown in Fig. 4. The figure shows that the

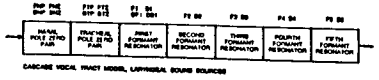

Fig. 4 Organization of the components of the KLSYN88 synthesizer for producing sounds with a source at the glottis.

bandwidths of the formants as well as their frequencies are controllable — an essential feature if a variety of voices are to be synthesized or if natural vowels and transitions are to be generated.

Theoretical and experimental work has suggested that a single pole-zero pair in the low- to midfrequency range is sufficient to provide natural-sounding nasal murmurs and nasal vowels with spectra that match the spectra of natural utterances [1], [3], [4], [5]. The frequencies and bandwidths of the nasal pole and zero are designated in the figure as FNP, FNZ, BNP, and BNZ. In the case of a nasal murmur, the sound output is from the nose, and the zero in the transfer function arises because of the side branch formed by the mouth cavity. For a nasal vowel, the output is from both the mouth opening and the nose, and the zero for an adult speaker is usually in the range 500 to 1500 Hz, depending on the vowel and the size of the velopharyngeal opening. When a nasal consonant is released into a vowel, the natural frequencies of the combined vocal- and nasal-tract configuration change continuously, although some of the frequencies (such as the lowest one) might change quite rapidly. The zero also changes continuously, but its motion can be extremely rapid, and may traverse 1000 Hz or more in a few milliseconds.

An example showing the trajectories of the poles and the zero for synthesis of a nasal consonant in intervocalic position is given in Fig. 5. In the first part of the initial vowel and after 20-odd ms of the final vowel, there is no nasalization, and the extra pole and zero can-

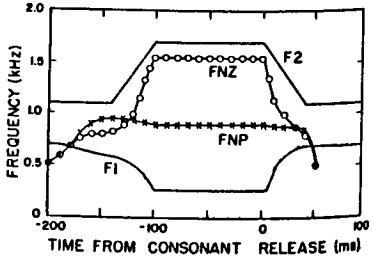

Fig. 5 Time course of several of the parameters used to synthesize the utterance /ana/.

cel each other. Creation of a velopharyngeal opening causes a separation of the pole and the zero. The velopharyngeal movement usually begins well in advance of the consonant closure, as shown by the separation of the pole and zero in the first vowel. As the constriction is formed and the output shifts from the mouth to the nose, there is a rather abrupt change in the zero. This rapid movement of the zero, together with the less extensive changes of the poles, causes a rather abrupt change in spectrum amplitude in the mid- and high-frequency range at the implosion and release of the consonant. A spectrogram of the synthesized sound is shown in Fig. 6. This spectrogram is similar to those observed in spoken syllables containing nasal consonants. The optimum trajectories for the additional pole and zero depend to some extent on the vowel, but theoretical principles as well as experimental data can help to guide the selection of these trajectories.

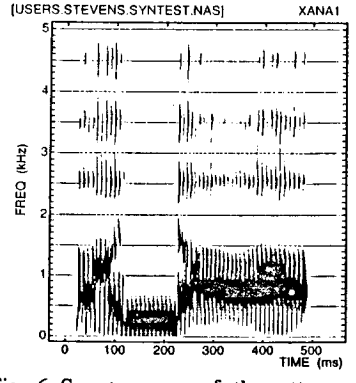

Fig. 6 Spectrogram of the utterance /ana/ synthesized using the parameters in Fig. 5.

When nasal consonants are synthesized by this procedure, the amplitude and spectrum of the glottal source remains fixed as the boundary between the consonant and the vowel is traversed. This is, of course, just what happens for human production of syllables of this kind. The abrupt change in spectrum amplitude at mid and high frequencies across a transition from a nasal consonant to a vowel occurs as a consequence of rapid changes in the characteristics of the transfer function and not as a consequence of a change in the source.

5. SYNTHESIS OF LIQUID CONSONANTS

The configuration with an additional pole-zero pair in the cascade branch makes it possible to synthesize liquid consonants in a natural way, as well as nasal consonants. As Fant showed in his early work [1], the transfer function of the vocal tract for a lateral consonant configuration is characterized by a zero or antiresonance, usually in the frequency range 2-3.5 kHz, as a consequence of the formation of a side branch in the acoustic path by the tongue blade. As the lateral consonant is released into a following vowel, this zero is annihilated or canceled by a pole. The elimination of the zero occurs rapidly, and results in an abrupt increase in spectrum amplitude in this higher frequency range.

Typical trajectories for the resonances and the antiresonance needed to synthesize a lateral consonant in a consonant-vowel syllable are given in Fig. 7. A retroflex consonant of the type that occurs in American English is also characterized by an additional pole-zero pair in the vocal-tract transfer function. The time course of the frequencies of this pole and zero, as well as of the other resonances, for a syllable like /ra/ is also given in Fig. 7. Not shown in the figure are bandwidth changes for some of the formants as the consonant is released from the

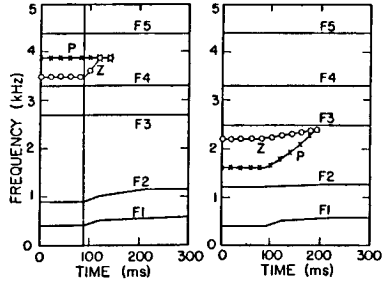

Fig. 7 Time course of the formant parameters and the additional pole-zero pair used to synthesize the syllables /la/ (left) and /ra/ (right).

constricted position. Acoustic analysis of synthesized syllables produced in this way shows spectral changes very similar to those in natural utterances. In particular, the discontinuity in spectrum amplitude at high frequencies at the release of the lateral consonant is achieved readily by the rapid movement of the zero in the transfer function.

6. GLOTTAL SOURCE AND TRACHEAL COUPLING

Perhaps the most significant improvement that Klatt made to the synthesizer configuration is related to the glottal source and the effect of the trachea. The details of the design of this source were given in a paper by Klatt and his daughter which appeared in the Journal of the Acoustical Society of America in February of 1990 [9]. The new synthesizer includes a controllable glottal source that is a modification of the source proposed by Fant, Liljencrants and Lin [2], and by Rosenberg [11]. The voicing source and the aspiration source are generated in the manner shown in Fig. 8. Not shown in the figure is the fact that the effects of the radiation characteristic have been folded into the source models, in effect yielding a waveform that is the time derivative of the output shown in this figure. The source controls are arranged so that adjustment of the open quotient OQ only affects the spectrum of the source at low frequencies, and has little influence on the high-frequency spectrum amplitude. The high-frequency amplitude is varied by manipulating the TL parameter. Numerically TL is the reduction (in decibels) in the spectrum amplitude of the source at 3 kHz, relative to the spectrum amplitude of a source with a simple 12 dB per octave slope at high frequencies.

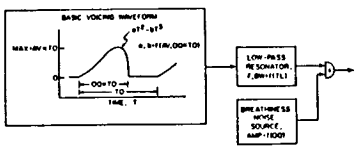

Fig. 8 Block diagram of the voicing source for the KLSYN88 formant synthesizer. The effects of the radiation characteristic have also been folded into the source models, resulting in a voicing source spectral output that falls off at about 6 dB/oct at high frequencies and an aspiration source spectrum that is essentially flat over the frequency range of interest.

Addition of an aspiration noise source to simulate turbulence noise in the vicinity of the glottis takes into account the spectrum of the turbulence noise and, when noise and voicing occur together, reflects the fact that the airflow, and hence the noise, are modulated. Furthermore, the turbulence noise source is, for the most part, a sound-pressure source, whereas the periodic glottal source is essentially a volume-velocity source. The spectral and temporal characteristics of the aspiration source in KLSYN88 are adjusted to take these factors into account, so that no further adjustment of the spectrum of the noise is necessary when the combined source forms the input to the cascade branch of the synthesizer.

An example of the smoothed spectrum of the synthesized vowel /a/, superimposed on the smooth spectrum for /h/ with the same formant frequencies and synthesized with the laryngeal source of Fig. 8 is shown in Fig. 9. The spectrum of /h/ in relation to that of the vowel shows substantial differences in amplitude at low frequencies but similar spectrum amplitudes in the F4 and F5 region. This comparison is consistent with acoustic data from natural speech as well as theoretical predictions [10], [12].

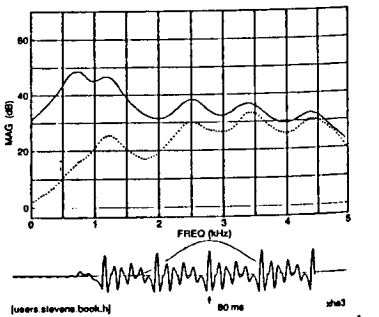

Fig. 9 Smoothed spectra of the sounds produced by the synthesizer for the vowel /a/ (solid line) and when the periodic glottal source is replaced by the aspiration source, with a widened first formant (dotted line).

Proper adjustment of the parameters OQ (open quotient), TL (high-frequency tilt), and AH (amplitude of aspiration noise) permits the generation of a glottal source with a spectrum that is a good approximation to the spectrum of the glottal source for almost any female or male speaker. Furthermore, within an utterance by a given speaker, these parameters can be varied as active laryngeal adjustments are made to produce voiceless obstruent consonants or prosodic changes within phrases and sentences. The parameters can also be modified as the laryngeal state reacts passively in response to manipulation of constrictions in the airway, for example during voiced obstruents and for sonorant consonants produced with a narrow constriction.

As we have seen in the cascade branch of the synthesizer in Fig. 4, two pole-zero pairs are included in addition to the series of poles for conventional synthesis of vowels. In addition to their use in the synthesis of nasals and laterals, one or both of these pole-zero pairs can be employed to simulate acoustic coupling to the trachea when the glottal opening is sufficiently large.

Proper positioning of a pole-zero pair introduces an additional peak in the spectrum at a relatively fixed frequency. The most prominent peak is usually the second tracheal resonance, which is in the frequency range 1400-1800 Hz for an adult.

To illustrate how the parameters of the glottal source can be manipulated to match the voices of different male and female speakers, we have attempted to match the spectra of selected vowels produced by several speakers, by manipulating the parameters of the synthesizer. In particular, we adjusted the frequency and amplitude of the glottal source and the formant frequencies and bandwidths to match the corresponding measured characteristics in the spoken vowels. We then manipulated the glottal parameters OQ and TL to provide a best match to the spectrum.

Examples of the spectra of the spoken vowels and of the best matching synthesized vowels are given in Fig. 10. These spectra illustrate quite diverse characteristics of the glottal source. For the female voice at the top, the parameter OQ was 70 percent, resulting in a prominent first harmonic. The first-formant bandwidth was wide (about 300 Hz), as might be expected with an increased average glottal width. On the other hand, the OQ value needed to match the male spectrum was 30 percent, with a first-formant bandwidth of 100 Hz. A slight high-frequency tilt (TL=2 dB) was necessary for this speaker.

For a number of voices it was possible to obtain a match to within 3-5 dB up to about 4 kHz by selecting values of the parameters OQ and TL. The values of OQ for different voices ranged from 30 to 70 percent, whereas TL was in the range 0 to 10 dB. For some voices, it was necessary to add a pole-zero pair to account for a minor spectral peak arising from acoustic coupling to the trachea. Some aspiration noise is routinely added to the glottal source during voiced intervals, and the amount of aspiration noise varies from speaker to speaker, presumably.

7. SYNTHESIS OF VOICED AND VOICELESS CONSONANTS: GLOTTAL SOURCE ADJUSTMENTS

As has been noted above, a speaker makes significant adjustments in the waveform of the glottal source during the production of various types of consonants, as well as in the syllabic nuclei over longer time intervals within a phrase or sentence. Dennis Klatt illustrated several of these types of adjustments in his paper published in 1990. Figure 11 shows a typical pattern of change of some of the glottal and other related parameters that are manipulated when a voiceless aspirated stop consonant in intervocalic position is synthesized. These parameters can best be interpreted in terms of the effects of the glottal spreading maneuver that occurs in conjunction with the supraglottal closing movement. The open glottis that assists the termination of voicing at the end of the first vowel

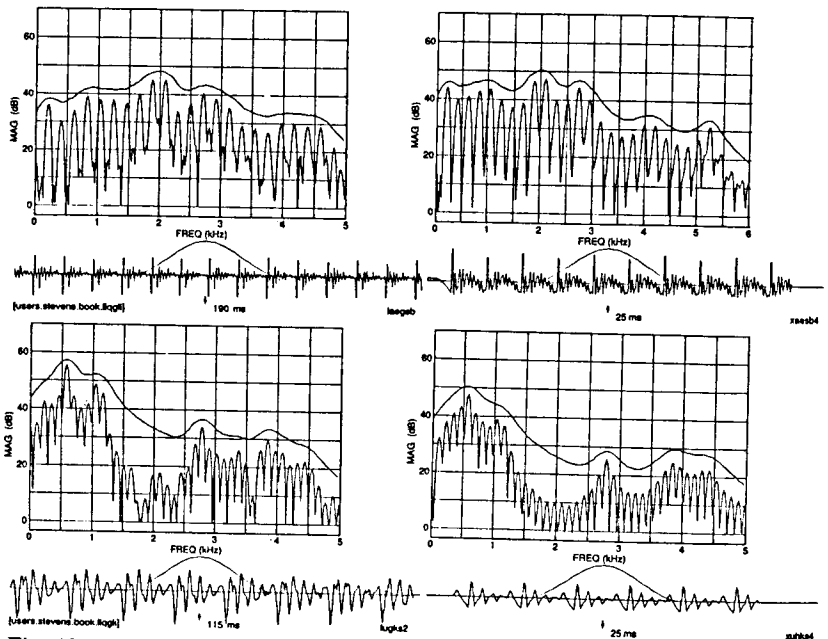

Fig. 10 The left panels show the spectra of two vowels produced by a female speaker (top) and a male speaker (bottom). The right panels are the spectra that are produced by the synthesizer when the glottal and bandwidth parameters are adjusted to give a best match. See text.

is reflected in an increased open quotient OQ, an increased high frequency spectral tilt TL, an increased amplitude of aspiration noise AH, and a widened bandwidth B1. The reverse occurs following the consonantal release preceding the onset of the second vowel. The transitions in formants F1 and F2 reflect the supraglottal movements toward and away from closure, and the burst of frication noise (identified by AF) also indicates the consonantal release. Thus a large number of synthesizer parameters need to be manipulated in order to provide an accurate acoustic representation of the glottal and supraglottal movements needed to produce the stop consonant.

8. TOWARDS REVISED RULES FOR SYNTHESIS FROM A PHONETIC INPUT

The few examples of synthesis of vowel-consonant and consonant-vowel sequences given here have indicated that if rules are to be formulated for synthesizing utterances from a phonetic input, these rules must specify the time variation of an extensive set of control parameters. We turn now to consider what these rules are trying to capture and how they might be organized.

In the case of vowels, the rules are relatively simple, and they specify the time course of a small number of formant frequencies --- usually just three formants. The glottal configuration specified for the vowels may change slowly depending on prosodic considerations. With little additional complexity, synthesis of the glides /w/ and /j/ can be specified in terms of formant movements. Since these glides involve a more constricted vocal-tract configuration, adjustment of the bandwidths of some formants may be necessary to account for the additional acoustic losses in the vocal tract, and some modification of the glottal vaveform (increased OQ and TL) may occur be-

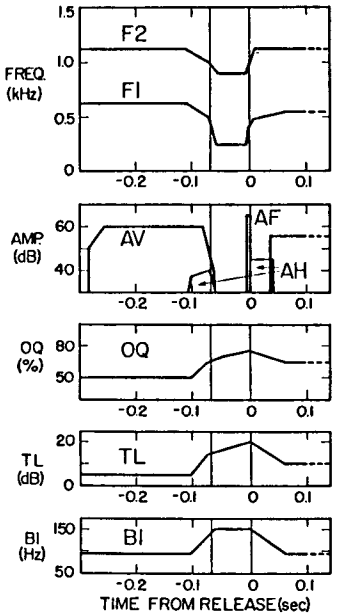

Fig. 11 Time course of several of the parameters used to synthesize the utterance /apə/. The vertical lines indicate the times of voicing offset and consonantal release.

cause of interaction of the glottal source and the vocal-tract acoustics. In the case of /h/, the spread glottal configuration is marked by appropriate adjustment of the parameters OQ, TL, AH, and B1.

For the consonantal segments the synthesis rules become substantially more complex. At least eight parameters need to be manipulated, for example, to synthesize the sound that occurs when a voiceless aspirated stop consonant is released into a vowel, as shown in Fig. 11. A similar large number of parameters is needed for nasal consonants and fricatives. Some of the parameters are related directly to the movement of the articulator that forms the consonantal constriction --- usually the lips, the tongue blade, or the tongue body. Other parameters are influenced primarily by adjustments of the laryngeal configuration, the velopharyngeal opening, or other articulators that are not directly involved in forming the constriction. There are, however, some interactions between these groups of parameters primarily through aerodynamic and acoustic processes. Thus, for example, the amplitude of the burst at the release of a stop consonant is determined in part by airflow that may be limited by the laryngeal configuration. Or the degree of abduction or adduction of the vocal folds can influence the frequencies and bandwidths of the lower formants. Or, the velopharyngeal opening can cause shifts in the natural frequencies of the vocal tract whose movements signal the place of articulation of a nasal consonant.

One approach to synthesis with the large array of parameters is to design a set of higher-level control parameters (HL parameters) that are related more closely to articulatory parameters than are the acoustically oriented parameters currently used to control KL-SYN88 (KL parameters) [13]. The arrangement is shown schematically in Fig. 12. The HL parameters specify articulatory dimensions such as size of velopharyngeal opening, area of a consonantal constriction, and glottal adduction-abduction. The lower-level KL parameters that control the synthesizer itself are derived from the HL parameters through a set of mapping relations. The mapping relations automatically incorporate the constraints that exist between the various KL parameters because of aerodynamic and acoustic interactions.

The generation of speech with the KL-SYN88 synthesizer requires that quantitative data and explicit models be developed in two areas of phonetics. One area is concerned with the constraints that the articulatory and aerodynamic systems impose on the sound. In terms of the diagram in Fig. 12, these are the mapping relations between the HL parameters and the KL parameters. The other area involves the temporal control of the articulatory processes, as these processes are manifested in the HL parameters.

Developing the mapping relations requires that theories and models of glottal vibration, aerodynamic noise generation and vocal-tract acoustics be re-

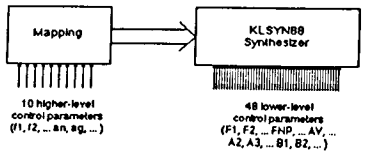

Fig. 12 Block diagram of a system for controlling the parameters of the KLSYN88 synthesizer by a reduced array of HL parameters. The KL parameters are derived from the HL parameters by a set of mapping relations.

fined. Examples of problems that must be addressed are: estimating the distribution of turbulence noise sources when there is a constriction in the vocal tract, determining the time course of onsets and offsets of vocal-fold vibration for voiceless consonants, and modelling the acoustic losses when there are consonantal constrictions.

Refinement of our understanding of articulatory control processes highlights the need for several types of data and models. Quantitative data must be obtained on rates of release and closure of articulators that form the primary consonantal constrictions for stops, fricatives and affricates. It is necessary to determine how articulatory parameters that are not directly involved in forming the consonantal constriction are timed in relation to the primary articulators. For example, is the control of these secondary articulators adjusted so that the acoustic evidence for the movements of these articulators is optimally represented in the sound?

Beyond these problems of controlling the synthesizer parameters to produce a representation of speech sounds in syllables, there are a variety of questions relating to timing and prosody of larger units. Klatt made important contributions here, with his extensive data and rules concerned with segmental durations [6] and his implementation of rules for generating fundamental frequency contours in phrases and sentences [8].

The existence of a speech synthesizer such as KLSYN88, with its ability to generate speech of high quality, focusses attention on these problems that are central to the study of phonetics --- problems relating to how individual sounds are produced, how different articulatory structures are coordinated, and how larger production units are organized. In this sense, the synthesis work of Klatt has not only contributed a body of knowledge to phonetics but has also provided a focus and a stimulus for future research.

9. ACKNOWLEDGEMENT

The preparation of this paper was supported in part by grant DC-00075 from the National Institutes of Health.

10. REFERENCES

[1] FANT, G. (1960), "Acoustic theory of speech production", The Hague: Mouton & Co.

[2] FANT, G., LILJENCRANTS, J., & LIN, Q.G. (1985), "A four-parameter model of glottal flow", Speech Transmission Labs. QPSR 4, Royal Institute of Technology, Stockholm, 1-13.

[3] FUJIMURA, O. (1960), "Spectra of nasalized vowels", Research Laboratory of Electronics QPR 62, Mass. Inst. of Technology, Cambridge, MA, 214-218.

[4] FUJIMURA, O. (1962), "Analysis of nasal consonants", J. Acoust. Soc. Am. 34, 1865-1875.

[5] HAWKINS, S.S. & STEVENS, K.N. (1985), "Acoustic and perceptual correlates of the non-nasal/nasal distinction for vowels", J. Acoust. Soc. Am. 77, 1560-1575.

[6] KLATT, D.H. (1979), "Synthesis by rule of segmental durations in English sentences", in Frontiers of speech communication research, edited by B. Lindblom and S. Ohman (Academic Press, New York), 287-300.

[7] KLATT, D.H. (1980), "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. 67, 971-995.

[8] KLATT, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Am. 82, 737-793.

[9] KLATT, D.H. & KLATT, L.C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am. 87, 820-857.

[10] MANUEL, S.Y. & STEVENS, K.N. (1989), "Acoustic properties of /h/", J. Acoust. Soc. Am. 86, Suppl. 1, S49.

[11] ROSENBERG, A. (1971),"Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am. 49, 583-590.

[12] STEVENS, K.N. (1990), "Noise at the glottis during speech production", J. Acoust. Soc. Am. 87, Suppl. 1, S121.

[13] STEVENS, K.N. & BICKLEY, C.A. (1991), "Constraints among parameters simplify control of Klatt formant synthesizer", J. Phonetics 19.