THE ROLE OF SYLLABLE STRUCTURE IN THE ACOUSTIC REALIZATIONS OF STOPS*

Mark A. Randolph and Victor W. Zue

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper examines the role of the syllable in the description of systematic acoustic-phonetic variations. We present results of an acoustic study based on over 5,000 stops collected from 1,000 sentences spoken by 100 talkers. Our results indicate that the acoustic properties of stops depend on the syllable locations in which they appear. On the basis of these results we propose a syllable-based rule framework in order to describe acoustic-phonetic variations in categorical as well as continuous terms. Implications to linguistic and speech recognition research are discussed.

INTRODUCTION

It is well known that the acoustic characteristics of speech sounds vary according to the context in which they appear. Traditionally, *systematic* acoustic variation has been described using context-sensitive *rewrite rules* of the form: $A \rightarrow B \ / \ C \_ D$, where elements A, C, and D correspond either to individual phonemes or classes of phonemes and element B corresponds to a specific phonetic realization [2]. As an example, rule (1) states that voiceless stop consonants are aspirated when followed by vowels.

$$\begin{Bmatrix} p \\ t \\ k \end{Bmatrix} \rightarrow \begin{Bmatrix} p^h \\ t^h \\ k^h \end{Bmatrix} / \_ V \qquad (1)$$

There are at least two disadvantages associated with such a rule description. First, it is awkward to describe the important role played by larger phonological units such as syllables or metrical feet. Second, it implicitly assumes that variations can be described in categorical terms, despite the fact that many acoustic changes are inherently continuous.

This paper proposes an alternative framework for describing acoustic-phonetic modifications. Central to this description is the notion of the syllable. We show how a rule framework based on the syllable may be augmented so as to describe contextual variations both concisely and accurately. We describe a set of acoustic studies focusing on the stop consonants in American English, and show that the proposed framework is well suited for interpreting the results. Finally, we describe the implications of the proposed framework for linguistic and speech recognition research.

THE SYLLABLE FRAMEWORK

The notion that phonological rules may be sensitive to syllable structure has been suggested by many linguists. Kahn [6], for example, argues that allophonic variation and phonotactic constraints can be described more effectively using a syllable-based phonological framework. Fujimura and Lovins [4] have provided articulatory data along with a summary of a number of acoustic-phonetic studies which provide concrete support for the syllable. Nakatani and Dukes [8] provide evidence from the perceptual domain. Their experiments indicate that the syllable-initial and syllable-final allophones of phonemes provide important perceptual cues for word juncture and that humans may rely on this kind of information for parsing phonetic sequences into words. While these studies provide compelling evidence in support of a syllable-based phonological representation, we are still in need of considerably more acoustic-phonetic data: quantitative results, derived from a large body of speech, showing that the surface acoustic realizations of phonetic units are governed by their positions within this unit.

In the next section, we show that if structured in the proper way, these results could be particularly relevant to the notion of a syllable hierarchy [3] [10], a structural description of the syllable in terms of an immediate constituent grammar. Linguists have found this hierarchical description important for the concise statement of phonotactic restrictions. As we will discuss later in this paper, this hierarchical representation also provides an effective means of incorporating the syllable into a description of acoustic-phonetic modifications.

THE CURRENT INVESTIGATION

We begin by describing the syllable template shown in Figure 1. We have used this template to label our experimental database and for the subsequent interpretation of our results. The form of this template closely resembles the syllable hierarchy proposed by Fudge [3]. We have modified his template by positing three affix positions and by providing labels for the *outer-onset*, *inner-onset*, *inner-coda*, and *outer-coda* positions. In addition, we have added an additional slot to the onset for the phoneme, /s/, which forms syllable-initial clusters with nasals, stops, and stop-semivowel sequences. The other terminal elements of this hierarchy are manner of articulation classes.

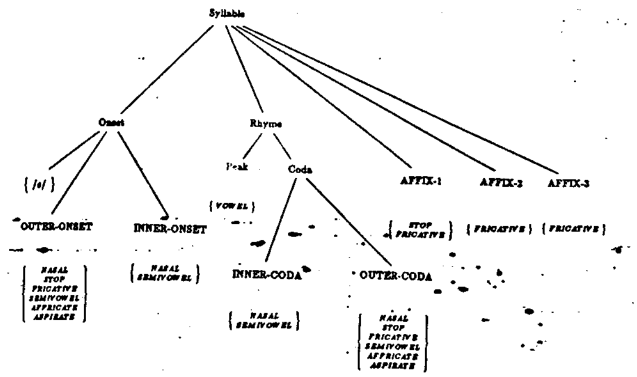

Figure 1: Syllable-constituent structure described in terms of broad phonetic categories

The Acoustic Study

Data for our acoustic study has been obtained from 1,000 sentences spoken by 100 talkers (50 male and 50 female). The corpus was the first five hundred of the well-known Harvard list of phonetically-balanced sentences. During recording, lists of ten sentences were read by one male and one female talker. For all the collected data, both phonemic and phonetic transcriptions were provided and aligned with the waveforms. In addition, syllable boundaries and lexical stress markers were inserted in the transcriptions. From this database, a sample of approximately 5,200 stops was extracted for the present set of experiments.

For each stop, we measured the closure duration and the release duration (VOT) separately. We also measured the durations of adjacent phonemes. From these measurements and transcriptions, we were able to determine whether a stop was released, unreleased, or deleted. We marked a stop as released if its release duration was greater than zero, unreleased if the release duration equalled zero, and deleted if the stop was present in the phonemic transcription, but absent in the phonetic. We should note that a stop was transcribed as unreleased if it could not be heard, and if a noticeable burst could not be observed from either the waveform or the spectrogram by the transcriber. In addition to duration measurements, we also computed several energy related parameters in order to infer the relative strength of a stop's release.

We are primarily interested in quantifying the effects of a stop's syllable position on these acoustic properties. However, we are also interested in understanding any possible influence of local phonetic context. In order to reduce the number of categories of local phonetic context to a reasonable size, we grouped the phonemes forming each stop's left and right context into seven equivalence categories corresponding roughly to manner of articulation. These categories are: Vowel (V), Semivowel (G), Nasal (N), Fricative (F), Stop (S), Affricate (A), and Aspirate (H).

Stops were categorized according to both local phonetic context and syllable position. Space limitations prohibit us from presenting data for all combinations of these two factors. Instead, we will present three examples from this larger pool of results. We will examine stops in two local phonetic environments, for each, we will examine the effect of syllable position on a stop's acoustic properties. In a third example, we examine the effect of post-vocalic voicing on vowel duration, also as a function of the stop's syllable position.

Results

Our first set of results compares intervocalic singleton stops in the outer-onset verses outer-coda positions. There were 668 outer-onset stops in this local phonetic environment, of which, 96% were released. In contrast, only 65% of 168 outer-coda stops were released. For singleton stops in the outer-onset, VOT is a reliable measure for voicing contrast. This can be seen from the histograms for voiced and voiceless stops shown in Figure 2. For syllable-final voiceless stops that were released (also shown in Figure 2), VOT is substantially reduced, such that there is considerable overlap of the distributions for outer-onset voiced stops and outer-coda voiceless stops.

The second example involves stop-semivowel sequences appearing between two vowels, i.e. the $V \_ GV$ context, where the stop is voiceless. In the outer-onset position (e.g., in the word sequence "gray train"), about 98% of the stops were released. On the other hand, only about 45% of the stops were released when they appeared in the outer-coda position (e.g. "great rain"). In Figure 3 we have plotted VOT versus the averaged total energy within the release for voiceless stops in both the outer-onset and outer-coda positions. We see that syllable-initial stops generally have releases that are both longer and stronger than their syllable-final counterparts.

Our final example concerns the effect of voicing of a stop on the duration of a preceding vowel. It is well known that the duration of the vowel is influenced by the voicing characteristic of the following consonant (e.g, the vowel in "bag" is longer than the vowel in "back") [9]. However, there seems to be evidence from our study that such influence is conditioned upon whether the vowel and stop belong to the same syllable. When the stop is in the outer-coda position, the preceding vowel is lengthened when the stop is voiced. However, the trend is reversed when the stop is in the onset of the following syllable. These results are summarized in Figure 4.

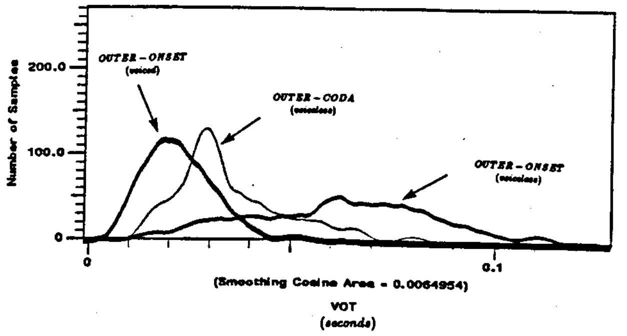

Figure 2: Influence of syllable position on the VOT of intervocalic, singleton stops

*This research was supported by ONR under contract N00014-82-K-0727, monitored through Naval Electronic Systems Command and AT&T Bell Laboratories Cooperative Research Fellowship Program

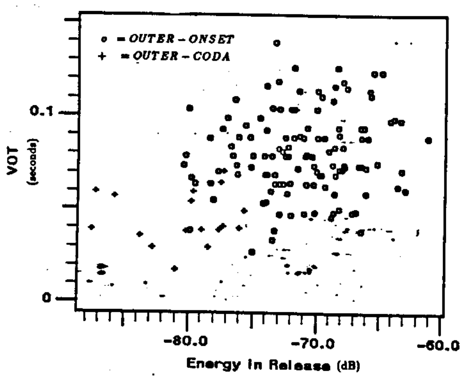

Figure 3: Influence of syllable position on voiceless stops in the V_GV context

DISCUSSION

From the results of our experiments, we may conclude that the acoustic characteristics of stop consonants depend on their positions within the syllable. However, our results also indicate that a more accurate description of these acoustic modifications may require an alternative rule framework in which acoustic information in the form of parameter values can be accommodated.

The Proposed Framework

The first aspect of our proposal is inspired by the work Church [1] and is motivated by principles of *information factoring*. The idea is to encode the description of a phoneme's contextual environment in terms of the syllable hierarchy. As a result, it becomes possible to replace a phonological grammar consisting of context sensitive rules by one which is context free. In general, context free grammars describe languages that are easier to parse, and in many cases, provide a more concise statement of phonological rules. For example, rather than inserting syllable boundary markers into a rule to describe the syllable positions for which stops are aspirated, one may describe these contextual environments more succinctly by restricting aspirated stops to particular slots within the template shown in Figure 1.

These new rules, however, since they ignore quantitative acoustic differences between phonemes appearing in the various syllable positions, still do not provide an adequate description of the facts. For example, aspirated stops can appear both in the outer-onset and outer-coda positions, but with differences in VOT that turn out to be important for determining the syllable structure of an utterance. The second aspect of our proposal is to augment this categorical representation with an acoustic description.

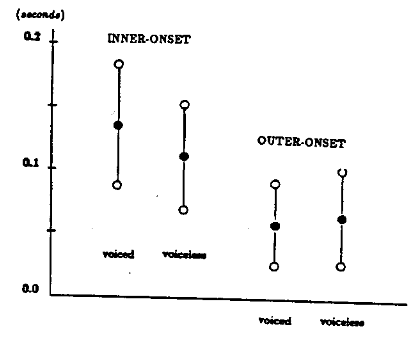

Figure 4: Influence of voicing and syllable position on preceding vowel duration

A more accurate mechanism would be to state these rules in the form of a conditional probability function such as the one shown in Equation (2).

$$p\left[\vec{A}|S,\sigma,\alpha,\beta\right] \tag{2}$$

The vector quantity $\vec{A}$ in this "rule" is a set of acoustic properties, some of which may be discrete (e.g., released, deleted, etc.), others may be continuous (e.g., VOT, the measured energy in release, etc.). The conditioning variables in this rule, or explanatory factors, are phonological in nature and reflect the phonemic identity of a segment and its phonological context. For example, the factor S in this rule may denote a particular phoneme (e.g., /p/, /t/, /k/, etc.) or a phoneme class (e.g., *STOP*, *FRICATIVE*, etc.), σ denotes S's syllable position (e.g., *outer-onset*, *inner-onset*, *peak*, etc.), and α and β specify the left and right context, respectively.

Since it attempts to describe the acoustic properties of phonemes directly, this rule framework bypasses an allophonic description of the speech waveform and therefore suggests a paradigm for research that is a hybrid of traditional phonetics and phonological methodologies [7]. The task involved in rule discovery is to seek a parsimonious combination of explanatory factors that best account for the acoustic-phonetic data. These steps would be carried out within the context of an acoustic study like the one described above.

Implications for Automatic Speech Recognition

The applicability of this probabilistic rule framework for automatic speech recognition may be readily seen by straightforward manipulations of the quantity shown in Equation (2). For example, given a particular syllable hypothesis, and a hypothesized local context, the *a posteriori* probability of a particular segment hypothesis is $p\left[S|\vec{A},\sigma,\alpha,\beta\right]$, and may be obtained using Bayes rule. In this function, the vector $\vec{A}$ denotes some appropriate set of acoustic parameters designed to identify S.

The quantity given in Equation (2) may also be useful for lexical retrieval. Church proposed a speech recognition framework in which a *narrow* phonetic transcription is parsed into syllables prior to lexical retrieval, using extrinsic allophonic variation as a means of constraint. The practical limitation of Church's approach is that it may not be possible to obtain such a detailed phonetic transcription from an acoustic front-end. However, a partial phonetic description of the speech signal in the form of a broad phonetic transcription consisting of a sequence manner categories, may be a more realistic alternative. This approach has been suggested by Huttenlocher and Zue [5] for the task of large vocabulary isolated word recognition.

Church's grammar would have to be rewritten, more along the lines of the syllable template shown in Figure 1. The direct consequence is a grammar which has a higher degree of ambiguity. Figure 5 shows the result of parsing the broad phonetic transcription of the phrase, "black lead." The output is provided in the form of a *syllable lattice*: a set of arcs (shown as rectangular boxes) spanning the input string. The arcs are labelled with the names of syllable constituents corresponding to what the parser has hypothesized: For this example, we see that the phoneme /k/ can be parsed as either the outer-coda of the first syllable or the outer-onset of the second. Such ambiguity arises because detailed phonetic information is no longer available. From Figure 3, however, we note that a voiceless stop in the outer-coda position will have reduced VOT and energy compared to its outer-onset counterparts. For this example, these attributes can be confirmed from the spectrogram in Figure 5.

Our approach to reducing the number of competing syllable hypothesis is to select a set of appropriately chosen acoustic attributes (e.g. VOT for stops) and to use the *a posteriori*, probability $p\left[\sigma|\vec{A},S,\alpha,\beta\right]$, to aid in disambiguating a parse. We believe that such a strategy offers the advantage of not requiring a detailed transcription to be available, while directly making use of acoustic measurements that are potentially more accurate. Efforts in implementing such a recognition strategy is currently under way.

SUMMARY

We have examined the role of syllable structure in the acoustic realizations of stop consonants in American English. The results of our acoustic study indicate that much of the apparent variability that a stop is subject to, may be explained in terms of its position within the syllabic unit. We have proposed a rule framework that is intended to capture this variability both concisely and accurately. Each rule in our framework is stated in the form of a conditional probability function. The conditioning variables (i.e., each rule's input) represent both the underlying phonemic identity of a segment and its phonological context. The rule's output is a description of its acoustic consequences. Finally, the relevancy of our proposal to linguistic and automatic speech recognition research was discussed.

REFERENCES

[1] Church, K. W., "Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints," Ph.D. Thesis, Massachusetts Institute of Technology, January 1983.

[2] Cohen P.S., and Mercer, P.L., "The phonological component of an Automatic Speech Recognition System," in *Speech Recognition*, R. Reddy, Ed., Academic Press, New York, pp. 275-320.

[3] Fudge, E.C., "Syllables, " *Journal of Linguistics*, Vol. 5, pp. 253-286.

[4] Fujimura, O. and Lovins, J., "Syllables as Concatenative Units," Indiana University Linguistics Club, 1982.

[5] Huttenlocher, D.P. and Zue, V.W., "A model of Lexical Access from Partial Phonetic Information," Proc. ICASSP, 1984

[6] Kahn, D., "Syllable-based Generalizations in English Phonology," Ph.D. Thesis, Department of Linguistics, Massachusetts Institute of Technology, September 1977.

[7] Liberman, M.Y., "In Favor of Some Uncommon Approaches to the Study of Speech," in *The Production of Speech*, MacNeilage, P.F., Ed., Springer-Verlag, New York, 1983.

[8] Nakatani, L., and Dukes, K.D., "Locus of Segmental Cues for Word Juncture," *J. Acoust. Soc. Am.*, Vol. 62, no. 3, pp. 714–719.

[9] House, A.S. and Fairbanks, G., "The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels," *J. Acoust. Soc. Am.*, Vol. 25, pp. 105–113.

[10] Selkirk, L.O, "The Syllable," in *The Structural of Phonological Representations*, Part II, Foris Publications, Dordrecht, Holland, pp. 337–383.

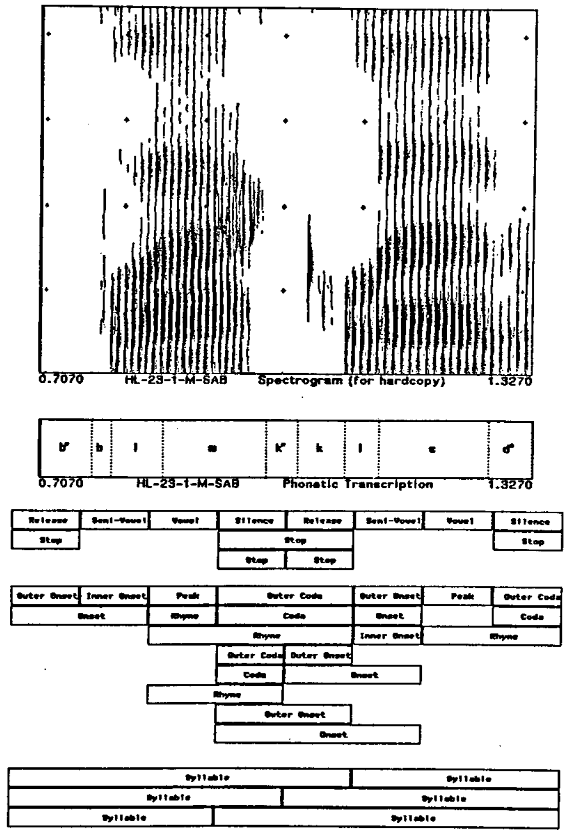

Figure 5: Syllable lattice generated from the broad syllable parser.