Studies of minimal speech sound units

Fant, G. and Lindblom, B.

journal: STL-QPSR
volume: 2
number: 2
year: 1961
pages: 001-011

http://www.speech.kth.se/qpsr
I. SPEECH ANALYSIS

A. STUDIES OF MINIMAL SPEECH SOUND UNITS

1. Theory of segmentation

The following study is an attempt to derive a specification system whereby spectrographic records of speech are described in terms of minimal units in the time-frequency-intensity space. The system should be detailed enough to retain any fact of interest for the correlation of speech wave signals with the production and perception of speech and with the units of speech messages as defined in linguistic theory. Yet it should be simple enough to be manageable and preferably permit various degrees of approximations for various purposes of analysis. In short, what is wanted is a system for condensed specifications of essentials of speech waves.

In several aspects the present approach is similar to that presented in earlier studies of distinctive features \(^{(1)}\)(\(^{(7)}\))(\(^{(8)}\)). Thus the theory of the phonetic realization of a phoneme as a bundle of distinctive features is still considered to be valid and powerful, reflecting basic properties of human speech. The view of speech as composed of one sound segment for each successive phoneme is an oversimplification of facts permissible only as an abstract interpretation in structural linguistics. One consequence of adopting such conventions in phonetic studies would be as follows. The extent of the formant transitions in the onset of a vowel serving as cues for the recognition of a previous consonant is completely conditioned by the F-pattern of the consonant segment proper and that of the vowel. Thus the consonant is specified by its energy distribution and its inherent F-pattern and consequently there would be only one segment per phoneme.

These were the concepts underlying the statements in Preliminaries to Speech Analysis \(^{(7)}\) though not so explicitly formulated. The present study is concerned with a more detailed analysis of the successive sound units and their observable characteristics. In the simple example above the initial transitional part of the vowel following the consonant would be treated as a separate sound segment contributing to the identification of both the consonant and the vowel.
and its duration if definable would be a desirable object of specification (Figs. 1-1, 1-3). Any position in time associated with a change in the type of primary acoustic source, e.g., voiced – voiceless, or of the spectrum type related to the opening or closing off of one part of the vocal tract cavity complex would be considered as a segment boundary in the present theory (Fig. 1-2). The number of sound segments of a sentence defined accordingly is greater than the number of phonemes.

A common example is a delayed onset of voicing in a normally voiced sound following an unvoiced consonant, e.g., in the [l] following a [k] in a [kl]-cluster. This asynchrony of phonation with articulation splits the [l] into two parts. A similar example leading to the specification of additional segments within vowels is the onset or termination of nasalization. The maximum subdivision of the speech wave interval conventionally assigned to a single phoneme occurs in fully developed stop consonants, displaying a sequence of occlusion, transient (transient response of the vocal cavities to a step-function in pressure at the onset of the explosion), short fricative, and aspiration (voiceless onset of a following vowel).

A single sound feature, e.g., voicing, nasalization, retroflexion, etc., often extends over several successive sound segments and even over the time domains of several successive phonemes. The several distinctive sound features constituting the acoustic realization of a phoneme are not always simultaneities within the same speech segment and generally have varying boundary locations and durations (Fig. 1-3).

The present specificational system is based on a primary division of sound features in two basic categories; those determining segment boundaries which will be called segment type features and those specifying the contents of speech segments beyond the categorization inherent in the type features which we call segment pattern features. Segment type features thus concentrate on the discontinuous aspects of speech, in phonetic terminology referred to as manner of production. The segment pattern features, on the other hand, correspond more to the place of articulation.
The specification of any segment thus starts out with a number of binary decisions as to the presence or absence of each of the type features. The analysis then proceeds with the segment pattern features which are measured in terms of more or less continuous functions the particular choice of which depends upon the segment type classification. These data may finally be interpreted in terms of the most probable place of articulation. In automatic speech recognition the data from a number of successive segments should enable a phonemic or allophonic identification.

The determination of segment boundaries and the identification of segment type features proceed in parallel during the initial stage of the analysis. Boundaries are generally associated with the absolute or relative appearance or discontinuation along the time scale of one or several type features. In those instances when a complete type feature identification of a piece of speech cannot be made from the first inspection of the spectrogram and supplementary speech records the natural procedure will be to determine the boundaries first and then complete the type feature identification.

It is quite evident that the segmentation process as outlined above is not unique. The number of segments recognized within a sentence and the exact place of some of the boundaries will to some extent depend on the investigator. Thus it may be difficult to decide on how many segments should be delimited within transitional intervals. The cues of the nasal feature, for instance, often creep gradually into the spectrogram which obscures the contrast between nasal and non-nasal segments. However, the final identification of the phonetic or phonemic contents of the speech would presumably not suffer from such variations. In the course of future work an attempt will be made to develop more strict rules for dealing with questionable cases.

2. Segment type features

The limited number of physiological functions within the speaking mechanism sets a natural limit to the number of different segment types. In conformity with the theory of distinctive features
of Jakobson et al. but with a broader purpose than phonemic analysis alone it is feasible to describe sound segment types by the presence of one or several basic elements which may be called segment type features. The following list represents a preliminary attempt at such a classification:

<table>
<thead>
<tr>
<th>Segment type features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source features</strong></td>
</tr>
<tr>
<td>1 voice</td>
</tr>
<tr>
<td>2 noise</td>
</tr>
<tr>
<td>3 transient</td>
</tr>
<tr>
<td><strong>Resonator features</strong></td>
</tr>
<tr>
<td>4 oclusive</td>
</tr>
<tr>
<td>5 fricative</td>
</tr>
<tr>
<td>6 lateral</td>
</tr>
<tr>
<td>7 nasal</td>
</tr>
<tr>
<td>8 vowellite</td>
</tr>
<tr>
<td>9 transitional xx</td>
</tr>
</tbody>
</table>

The source features pertain to the type of acoustic sound source within the vocal tract and thus represent the phonation in a wider sense. The noise feature refers to random noise from turbulent airflow through narrow passages and past sharp obstacles. By transient source is meant a single shock excitation of the vocal cavities due to a sudden release of an over-pressure or a sudden checking of an airflow at any obstruction in the vocal cavities, the vocal cords included. In this sense voicing is identical with a sequence of quasi-periodical transients.

In a broad-band spectrogram of a voiced stop the transient is seen as an additional vertical striation which is non-synchronous with the pitch pulses. The transient of the voiced stop is generally followed by a very weak noise segment. In unvoiced stops the transient is typically followed by a noise segment of vowellike or fricative type, the vowellike part, if present, starting later than the fricative part.

x) A fourth source feature labeled silence is convenient for practical specificational work in so far as it replaces three negations, i.e., minus voice, minus noise, and minus transient. It is, however, completely redundant and thus not included in the list above.

xx) An intermediate degree of feature 9 labeled glide has been adopted in some of the preliminary studies (Fig. I-1). Although phonemic motivations exist for this extra distinction it does not seem practical to maintain it in the acoustic analysis.
In one and the same sound segment it is possible to find any combination of the source features 1, 2, and 3.

The resonator features may be described at the level of speech production as follows:

4. **Occlusive**  
   Complete or relative closure in the mouth or in the pharynx;

5. **Fricative**  
   Turbulent airflow caused by a supraglottal obstruction;

6. **Lateral**  
   Central closure combined with lateral opening in the mouth cavity;

7. **Nasal**  
   Nasal passages connected to the rest of the vocal system owing to a lowered velum;

8. **Vowel-like**  
   Free passage for the air stream through the pharynx and the mouth cavities;

9. **Transitional**  
   The articulators moving at a high speed within the segment. A "glide" would be defined by the articulators moving at a moderate speed within the segment.

The speech wave correlates of the resonator features may be described as follows:

4. **Occlusive**  
   The occlusive sound segment has a lower frequency $F_1$ and lower intensity levels of $F_2$ and upper formants than adjacent non-occlusive segments. The spectrum of a voiced non-nasal occlusive is dominated by a formant $F_1$ of a very low frequency $F_1$ (the voice bar). However, with considerable high-frequency pre-emphasis it may be possible to detect $F_2$ and $F_3$.

5. **Fricative**  
   The spectra of fricatives display a relatively prominent high-frequency noise energy. A necessary requirement for the fricative feature is thus the noise energy which is largely contained in formants from cavities in front of the articulatory narrowing. Spectra of voiced fricatives can in addition display the whole F-pattern up to $F_4$ or at least $F_1$.  


6. **Lateral**

Sound segments of lateral articulation produced with a voice source possess the vowel-like feature except for a selective intensity reduction in the region of the second and/or the third formant due to the first zero of the shunting mouth cavity behind the tongue. An additional high frequency formant is generally seen. The oral break provides a typical discontinuity in the connection to a following vowel. The lateral sound segment is generally, but not always, of lower frequency $F_1$ than a following or preceding vowel.

7. **Nasal**

A voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which $F_2$ is weak or absent. A formant at approximately 250 c/s dominates the spectrum, but several weaker high-frequency formants (not always seen in spectrograms) occur, one typically at 2200 c/s. These higher formants are generally weaker than for laterals. The bandwidths of nasal formants are generally larger than in vowel-like sounds. Voiced vowel-like nasal sounds (nasalized vowels) possess the nasal characteristics as a distortion superposed on the vowel spectrum. Typical nasalization cues are the addition of the first nasal resonance in the region below the first formant of the vowel-like sound and a simultaneous weakening and shift up in frequency of the first formant $F_1$. A relative reduction of $F_3$ is often apparent.

8. **Vowelike**

The F-pattern formants are clearly visible in the spectrogram. In the case of voiced or unvoiced vowel-like sounds produced with a glottal source it is required that at least $F_1$ and $F_2$ be detectable. $F_3$ should also be seen providing $F_1$ and $F_2$ are not located at their extreme low frequency limits. A specific feature of sounds produced with a glottal source is that the relative formant levels are highly predictable from the particular F-pattern, i.e., from the formant frequency locations. Vowel-like noise sounds produced from a supraglottal source possess a rather weak first formant, $F_1$. This is especially the case with [h]-segments.
produced with the tongue articulation of a high front vowel. Unless the fricative feature is superimposed there should not occur a prominent high-frequency noise area in the sound spectrum.

9. Transitional

The spectrum changes at a relatively fast rate in the segment. The first part of a vowel following a voiced stop or nasal is characterized by a rapid change in at least one formant frequency, e.g., $F_1$. The transitional sound segment ends where the major part of the formant transition is completed. The possible intermediated degree, glide, would be characterized by the spectrum changing at a relatively slow rate but faster than for a mere combination of two vowel phonemes. Variants of [r] [l] [j] [w] sounds occur as glides.

The possible co-occurrence of resonator features is partially evident from articulatory phonetics. However, the inventory of all the possible combinations has not yet been established. A few combinations of operational importance could be mentioned here. The fricative feature would be attributed to a sound segment corresponding to what is traditionally called a voiced fricative only if motivated by the presence of significant high-frequency noise. The occlusive and vowel-like features are not intended to be mutually exclusive. The addition of the vowel-like feature to the occlusive feature would normally differentiate voiced fricatives from voiced stop gaps. An exception is voiced stops in unstressed positions where the closure may be incomplete in which case the vowel-like feature has to be added to the description in view of the apparent formant pattern. The open phase of a rolled [r] would be labeled vowel-like non-occlusive and the closed phase would always be occlusive, with or without the vowel-like attribute.

---

x) The distinction between /b/, /v/, /w/ lies in part in the degree of contrast with regard to adjacent non-occlusive segments. The contrast is both a matter of speed of formant frequency variations, the differences in $F_1$, and the associated differences in the intensity levels of the upper formants. The critical values of these parameters are a function of the degree of stress.
There is a very close connection between the inventory of type features specified above and the parameters of the OVE II synthesis. The major exception is that the laterals are not treated separately from vowels in the synthesis and that the transient source either may be omitted from the synthesis or produced as a very short noise segment. In synthesis the occlusive feature is incorporated by the choice of a relatively low frequency $F_1$ of a segment and the transitional feature is a matter of the speed of variation of the $F$-pattern.

In synthesis the nasalization of a vowel adjacent to a nasal consonant adds not only to the naturalness of the nasal but also to the identification of the nasal element.

3. Segment pattern features

Most of the pattern features may be quantized in terms of an arbitrary number of standard articulatory positions or in terms of the acoustic parameter values allowing such identifications. A final identification of the phonemes of a piece of speech sometimes requires a few acoustic data such as intensity, duration, and speech of transitions within segments supplementing the place of articulation data. In other words, not all of the pattern features are related to the place of articulation. The articulatory patterns may be described as follows:

<table>
<thead>
<tr>
<th>Articulation</th>
<th>Speech wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Tongue fronted</td>
<td>$F_2 - F_1$ large.</td>
</tr>
<tr>
<td>a) Prepalatal position</td>
<td>$F_2$ high, $F_3$ maximally high.</td>
</tr>
<tr>
<td>b) Midpalatal position</td>
<td>$F_2$ maximally high and close to $F_3$.</td>
</tr>
<tr>
<td>12. Tongue retracted</td>
<td>$F_2 - F_1$ small. $F_1$ comparatively high.</td>
</tr>
<tr>
<td>13. Mouth-opening (including tongue section and lips narrow)</td>
<td>$F_1$ low.</td>
</tr>
<tr>
<td>14. Lips relatively close and protruded (small lip-opening area)</td>
<td>$F_1 + F_2 + F_3$ lower than with larger lip-opening and the same tongue articulation. A progressing lip closure alone causes a decrease in each of $F_1$, $F_2$, and $F_3$ but with varying amounts depending on the particular tongue position. The effect on $F_3$ is pronounced in case of prepalatal tongue positions.</td>
</tr>
</tbody>
</table>
15. Retroflex modification
   a) Alveolar articulation \( F_4 \) low and close to \( F_3 \).
   b) Palatal articulation \( F_3 \) low and close to \( F_2 \).

16. Bilabial or labiodental closure
   \( F_2 \) in the region of approximately 500 - 1500 c/s depending on the tongue location of the associated vowel or vowel-like segment. A palatal tongue position favors a high \( F_2 \). The noise spectrum of the fricative [ʃ] is essentially flat and of low intensity.

17. Interdental articulation
   \( F_2 \) 1400 - 1800 c/s. Fricative noise of [θ] much weaker than for [s] and a more continuous spectrum. Center of gravity is higher than for the labiodental fricative [ʃ].

18. Dental or prealveolar articulation
   \( F_2 \) in the region of 1400 - 1800 c/s, \( F_3 \) high. Fricative noise strong. The main part of the [θ]-energy is above 4000 c/s. This cutoff frequency is lower for alveolars than for dentals.

19. a) Palatal retroflex articulation
   \( F_3 \) low. The fricative noise of [ʃ] is of high intensity and is carried by \( F_3 \) and \( F_4 \).
   b) Palatal articulation with tip of tongue down
   \( F_2 \) and \( F_3 \) high. Strong noise centered on \( F_3 \) and \( F_4 \) and also on \( F_2 \) providing the tongue pass is sufficiently wide. The lower frequency limit of [θ]-noise is higher than for retroflex sounds.

20. Velar and pharyngeal articulation
   \( F_2 \) medium or low. A large part of the fricative noise is carried by \( F_2 \). The F-pattern except \( F_1 \) is clearly visible.

21. Glottal source
   The entire F-pattern including \( F_1 \) is visible.

The existence of complex articulations should be kept in mind. The most apparent example referred to above is the freedom of the tongue to take any position during lip closure which makes the F-pattern of labials variable. In dentals the back of the tongue is partially free to approach the back wall of pharynx which lowers \( F_2 \).
and increases $F_1$. This is the case of the "dark" l. The articulatory contrast between a wide unobstructed and a narrow divided pharynx, resulting in a high versus a low $F_2$, is the counterpart of the hard/soft distinction in Russian consonants.

4. Application of segmentation techniques

The segmentation methods have been applied to the utterance "Santa Claus". Segment boundaries have been traced in the spectrogram of Fig. I-1. The associations of segments with type features are presented in Fig. I-2 and with phonemes in Fig. I-3. No attempt has been made to state the relative importance of the various segments.

It is not claimed that the processing of this utterance is unambiguous. The transcription, i.e., choice of phonetic symbols under the spectrogram, is made according to what is heard although it is always tempting to influence the choice by a physiological interpretation of the spectrogram. Thus from a speech production point of view the /t/ could have been transcribed as a nasalized [d] or [n] alone. The impression of the final [z] and not [s] is probably backed up by the superimposed vocalic element (aspiration) of a second formant which is more prominent than what could be expected from an [s]-sound.

The co-articulation effects are apparent, e.g., in the extension of the nasal segment feature in the form of a divided $F_1$ over the entire [æ] and half the [θ]. It is not quite clear how much of the [k] that is co-articulated with the [l]. According to Truby (12) it is common to find the entire [k] of a [kl]-cluster to be lateralized.

There remains much work to be done on the comparative study of speech production and speech waves in human speech and synthetic speech before we can make full use of the potentialities of acoustic speech records as a documentation of the speech production events.

C. Fant, B. Lindblom

Fig. 1-1  Segmentation of a broad-band spectrographic record.
Text: "Santa Claus".
Fig. I-2 Classification of the acoustic segments of the utterance of Fig. I-1 in terms of segment type features.

Fig. I-3 Estimated phoneme-segment dependency of the utterance of Fig. I-1


