5e CONGRES INTERNATIONAL D'ACOUSTIQUE
LIEGE 7-14 SEPTEMBRE 1965

74 A51

Voice Onset Time in Stop Consonants: Acoustic Analysis and Synthesis

Arthur S. Abramson* and Leigh Lisker†

Haskins Laboratories, New York

* Also, Queens College, the City University of New York
† Also, the University of Pennsylvania

In nearly all known languages there are consonant phonemes articulated at a given point in the vocal tract that are minimally distinguished from each other by the nature of the sound source, whether quasi-periodic or aperiodic. This source difference seems to underlie a number of allegedly independent dimensions that are commonly used in the phonetic literature. Our research is concerned with the role of this source feature in the production and perception of stop consonants, a class of speech sounds whose characteristic feature is an interval during which the breath stream is completely blocked within the oral cavity. Currently we are looking at initial prevocalic stops, where the air pressure built up behind the oral closure is released more or less explosively as the vocal tract moves toward a configuration appropriate to some vowel.

Although much physiological work remains to be done, we may make certain inferences as to how the action of the larynx in conjunction with supraglottal events generates the acoustic features that characterize initial stops. We may suppose that the initiation of speech activity involves ordinarily the development of a subglottal overpressure and a consequent movement of air through the glottis into the supraglottal cavities. If, to begin with, the glottis is open, it offers little hindrance to the airflow. At some point during normally phonated speech, however, the speaker closes down the glottis, and, given a suitable balance of aerodynamic forces and muscular tension, the vocal folds will begin to vibrate. This shift in mode of laryngeal operation occurs more or less in step with a change in supraglottal articulation, from the closure phase of the stop to the progressively more open tract of the following vowel. The acoustic consequences of this combination of laryngeal and oral gestures depend very much on their relative timing. Laryngeal vibration provides the periodic or quasi-periodic carrier that we call voicing. Voicing yields harmonic excitation of a low-frequency band during closure, and of the full formant pattern after release of the stop. Should the onset of voicing be delayed until some time after the release, however, there will be an interval between release and voicing onset when the relatively unimpeded air rushing through the glottis will provide the turbulent excitation of a voiceless carrier commonly called aspiration. This aspiration phase is accompanied by considerable attenuation of the first formant, an effect presumably to be ascribed to the presence of the tracheal tube below the open glottis. Finally, the intensity of the burst, that is, the transient shock excitation of the oral cavity upon release of the stop, may vary depending on the pressures developed behind the stop closure, where such pressures will in turn be affected by the phasing of laryngeal closure. Thus it seems reasonable to us to suppose that all these acoustic features, despite their physical dissimilarities, can be ascribed ultimately to actions of the laryngeal mechanisms. We shall present data on the voice onset timing (VOT) of stops in

three languages --English, Spanish and Thai-- with the responses of speakers of those languages to a set of synthetic speech stimuli that varies in VOT values over a suitably broad range.

Among the acoustic features which reflect the particular sound source involved in the production of initial stops we have chosen to measure the time of onset of glottal periodicity relative to the burst which marks the release of stop closure. This choice was dictated by the fact that the onset of glottal pulses, seen as vertical striations in wide-band spectrograms, is the clearest sign that the glottis has shifted from the open to a closed position. Moreover, in a recent study (WORD, 20.3, 1964, 384-422) in which such VOT data are reported for word-initial stops in English and ten other languages, we have shown that this single measure serves as a very effective basis for assigning a stop to its proper linguistic category, despite variation from language to language both in the number of stop phonemes and in the particular phonetic features said to distinguish them. The three languages reported on here were chosen as fairly representative of the eleven previously studied. English and Spanish are both described as having voiced and voiceless stops, but the categories in the two languages reportedly differ in other phonetic features. The third language, Thai, has three categories of stops, described as voiced, voiceless unaspirated and voiceless aspirated.

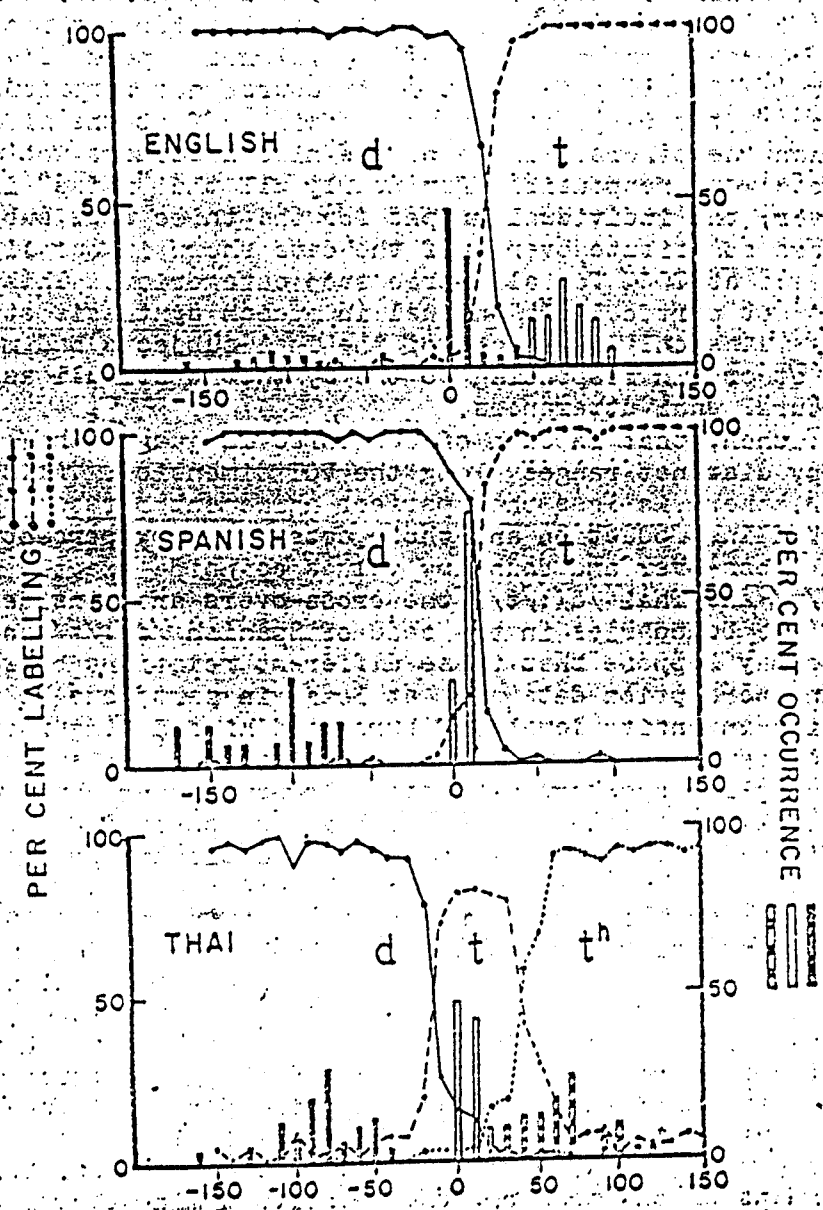
To make the stimuli needed for our perception experiments we used the Haskins Laboratories' speech synthesizer called Voback, an optical control system which accepts spectrographic patterns as inputs to the synthesizer end of a 19-channel vocoder. The basic pattern used in making our test stimuli consisted of a 15-msec high-frequency burst and a 150-msec three-formant transition appropriate to an initial apico-dental stop, followed by 400 msec of steady-state formants having the frequencies of an [a]-vowel. The onset of voice was simulated by simultaneously turning on the buzz source and the lowest spectral channel. The timing of this voice onset was varied in 10-msec steps from a value of -150 msec ("voicing lead") to a value of +150 ("voicing lag"). For cases of voicing lead the signal begins with an interval of low-frequency buzz, while for voicing lag the first formant is suppressed and the upper formants as well as the burst are excited by hiss until the point of buzz onset. The intensity level of the hiss source and the spectral characteristics of the burst were fixed, on the basis of preliminary testing, at values representing a compromise between the low-intensity burst normal to stops with voicing lead and the high-intensity burst and hiss characteristic of stops with long voicing lag. The 31 test stimuli made in this manner were recorded in several randomized orders for presentation to native speakers of American English, Puerto Rican Spanish and standard Thai, who were asked to label each stimulus in terms of their own phonemic categories.

Measurement and labeling response data for the apical stop categories of each language are presented for direct comparison in the three displays of our figure. The bars represent frequency distributions of measured VOT values, while the curves give the distributions of subjects' identifications of the synthetic speech stimuli as functions of our analogue VOT values.

The upper part of the figure shows our data for the English /d/ and /t/ categories. There is a very simple relation between the production and perception data: the 50% cross-over point for the /d/ and /t/ labeling curves is at +25 msec, and thus coincides neatly

with the boundary between the two non-overlapping ranges of VOT values for productions. It may be noted that the distributions for the two categories are quite different; VOT values for /t/ show a Gaussian distribution about a mean at +70 msec, but /d/ exhibits two discontinuous ranges that reflect the existence of two variants of this phoneme in initial position.

The middle display in the figure shows the comparable data for Spanish. VOT values for the two stop categories of this language are generally to the left as compared to the English stops: Spanish /t/ shows a modal VOT value at +10 and thus falls within the range of English /d/ rather than English /t/, while Spanish /d/ always shows long voicing lead. This leftward "shift", however, is not



reflected in the Spanish responses to our test stimuli; the perceptual cross-over does not match the boundary between the two ranges of measured VOT values, but somewhat surprisingly occurs just to the right of the /t/-range. Although the VOT dimension, then, provides sufficient information for perceptual sorting, it seems that the stimuli are not quite right for Spanish listeners. Even so, the Spanish cross-over is 10 msec to the left of the English cross-over, and enough responses by both groups of listeners were obtained to convince us that this difference is significant.

The Thai data in the bottom graph demonstrate very clearly the existence of three categories with respect to the VOT dimension. The perceptual cross-over between /d/ and /t/ falls within the boundary zone which separates the two ranges of VOT values. On the other hand, the cross-over at about +40 between /t/ and /t^h/ is displaced somewhat to the right of the boundary at +20, a value at which there is some slight overlap between the VOT ranges for these categories. It should also be observed that although the labeling curves obviously reflect three distinct perceptual categories, they do not reach values of 100%. Indeed, the middle category reaches a peak of no more than 84% at the +10 msec point. It is of course not surprising that irregularities in the subjects' responses will affect the middle category more than the others. Three of the nine listeners who served as our subjects were responsible for nearly all the "noise" in the data. Moreover, the individual graphs for the three best subjects, whose responses constitute over half the data shown in our figure, exhibit plateaus at 100% for all three stop categories. This state of affairs is not particularly unusual in speech synthesis studies; it presumably means only that some of the Thai subjects had more trouble than others in responding to the synthetic stimuli as though they were natural Thai utterances.

We find, then, that in each of our three languages the stop categories occupy distinct ranges along the VOT dimension. At the same time, however, there is less than perfect agreement between the ranges observed in production and those determined by perceptual testing. While cross-over and boundary values coincide exactly for English /d/:/t/ and Thai /d/:/t/, the cross-overs are considerably to the right of the boundaries in the case of Spanish /d/:/t/ and Thai /t/:/t^h/. We may suppose that these different relations between our production and perception data are, at least in part, due to the fact that in setting intensity levels of burst and hiss during the exploratory phase of synthesis, we were influenced to some extent by the reactions of English speakers. Conceivably some other choice or choices of burst and hiss intensity would bring cross-overs and boundaries for Spanish /d/:/t/ and Thai /t/:/t^h/ into closer agreement; this might or might not preserve the present close agreement in the English data. In addition to supposing that differences in level of burst and hiss intensity may serve to supplement the information provided by VOT, we might also infer from our data that Thai speakers are particularly sensitive to the presence of buzz before the burst. This sensitivity is, perhaps, to be referred to the fact that Thai speakers must differentiate three categories along a physical dimension that English and Spanish speakers utilize for only two.

To summarize, this study shows that stops in languages as diverse phonetically as English, Spanish and Thai can be identified acoustically by means of one measure, the relative timing of voice onset and release; comparison with listeners' responses to synthetic stimuli confirms the perceptual significance of this physical measure. Some of our data suggest experimenting with acoustic manifestations of glottal activity not yet studied.