SOME SUGGESTIONS FOR LANGUAGE TEACHING METHODS ARISING FROM RESEARCH ON THE ACOUSTIC ANALYSIS AND SYNTHESIS OF SPEECH

by

PIERRE C. DELATTRE,
University of Pennsylvania

FRANKLIN S. COOPER,
Haskins Laboratories, New York

ALVIN M. LIBERMAN,
University of Connecticut

Reprinted from: "Report of the Third Annual Round Table Meeting on Linguistics and Language Teaching," The Institute of Languages and Linguistics, Washington, D. C., *Monograph Series* Number 2, pp. 31-47, September, 1952.

In going over the research problems that have been treated by acoustical analysis and synthesis at the Haskins Laboratories in the course of the last few years, we have found several instances in which the results should have some value in the devising of teaching methods. But before turning to a discussion of the possible applications of this and related work, it perhaps would be well to review briefly the work itself and the general sort of results that are being obtained.

In acoustic phonetics, one of the problems is to find an appropriate way to represent and describe the acoustic events. The sound spectrograph is becoming a generally accepted tool for this purpose; it not only immobilizes the sound so that it may be studied at leisure but it provides also a "picture" which seems reasonably comprehensible to the eye. By studying a variety of spectrograms, one can begin to see the relation between the acoustic events and the perceived sounds of normal speech. Also, one can distort the pronunciation intentionally and observe the change in the spectrogram. But there are obvious and very narrow limits to the variations in sound which can be produced by the human voice, and consequently, the spectrograms will not answer all the questions which we should like to ask.

Spectrograms will usually exhibit several distinct features for any given phonetic unit or combination of units, and in that sense, the information which comes from the spectrogram is ambiguous with reference to the relation between acoustic stimulus and perception. For example, one looks at spectrograms of a given speech unit, and wonders which of the regularly occurring features are redundant for the identification of the sound, and which are not.

The most casual inspection of the stop consonants raises several questions: for example, does the recognition of [k] at the beginning of a word depend upon the characteristics of the initial burst of noise, or as the spectrogram seems to suggest, is recognition determined by a larger pattern which includes the [k] explosion and also the following vowel?

Spectrograms of connected speech show much more formant movement than steady state, and we should like to know about

the role of formant movement in the perception of the speech stream. For example, in the case of [l], the spectrogram typically shows that the formants glide and then reach a steady state. To what extent does the identification of the [l] sound depend on the steady state, and to what extent on the direction and rate of change of the formants? If the dynamic characteristics of the sound are involved, what then is the essential pattern on which perception depends?

To answer these questions, and many others, it is convenient, if not indeed necessary, to experiment with speech—that is, to make controlled modifications of the sound, and then to evaluate the effects of these modifications on the sound *as heard.* For this purpose, an instrument, called a pattern playback, was developed at Haskins Laboratories. The experimental work which we shall describe for you was done with this instrument.

In principle, the playback is somewhat like a player-piano except that a spectrogram replaces the perforated piano roll and the individual sounds are pure tones rather than harmonic-rich notes from a piano. Briefly, the playback generates 50 harmonic tones, 120 cycles apart, from 120 to 6000 cycles, in the form of beams of light modulated by a tone wheel. If the spectrogram is drawn with white paint and is made to pass under the modulated light, each painted portion of the spectrogram will reflect light and cause the corresponding harmonic to be heard when that light is converted into sound by means of a phototube. The principal advantage of such a machine is that it enables one to experiment with the dynamic aspect of speech sounds—that is, the rapid changes of formant frequencies in time—though it can also be used to deal with steady-state sounds.

We have found that this method provides a very convenient basis for experimenting with the perception of speech—that is, for making a great variety of changes in the acoustic stimuli and then determining the effects of these changes on the sound as it is heard. The method has been used to determine the acoustic correlates of nasality in French nasal vowels, to find a reasonably satisfactory way of producing the cardinal vowels with two formants only, and, in several exploratory

studies, to find the effects on intelligibility of the omission and modification of various aspects of the speech pattern, including alterations in the rate of change of the various formants.

Let me trace for you, now, the general path along which this research has proceeded, with some recordings to illustrate the kinds of sounds with which we are experimenting.

One of the very first things we tried was, of course, connected speech played back directly from spectrograms. Here is a spectrogram showing nonsense sentences.

(An original spectrogram on photographic film of three standard test sentences was shown, and recording was played of the three sentences as converted into sound by the playback.)

In order to secure greater flexibility in manipulating speech sounds, we paint by hand spectrograms such as this one, which shows the same sentences in much simplified form.

(A hand painted spectrogram was shown.)

Here is a recording which will let you compare the synthetic speech from an original spectrogram, from a detailed painting made from this spectrogram, and from a much simplified painted spectrogram. The phrase is "Never kill a snake."

(The three recordings of Fig. 1 were shown, and recordings of them played.)

Here, also, are several more of the test sentences, all synthesized from painted tapes like the one we have shown.

(Recordings of four sentences in simplified form are played as spoken by the playback.)

FIG. 1. Three versions of the phrase "Never kill a snake." TOP: A spectrogram containing full information about the spoken phrase, i.e., a spectrographic analysis of normal speech. MIDDLE: A painted spectrogram in which an attempt was made to include much of the detail of the photographic version. BOTTOM: A painted spectrogram which was considerably simplified and schematized, but with little loss of intelligibility.

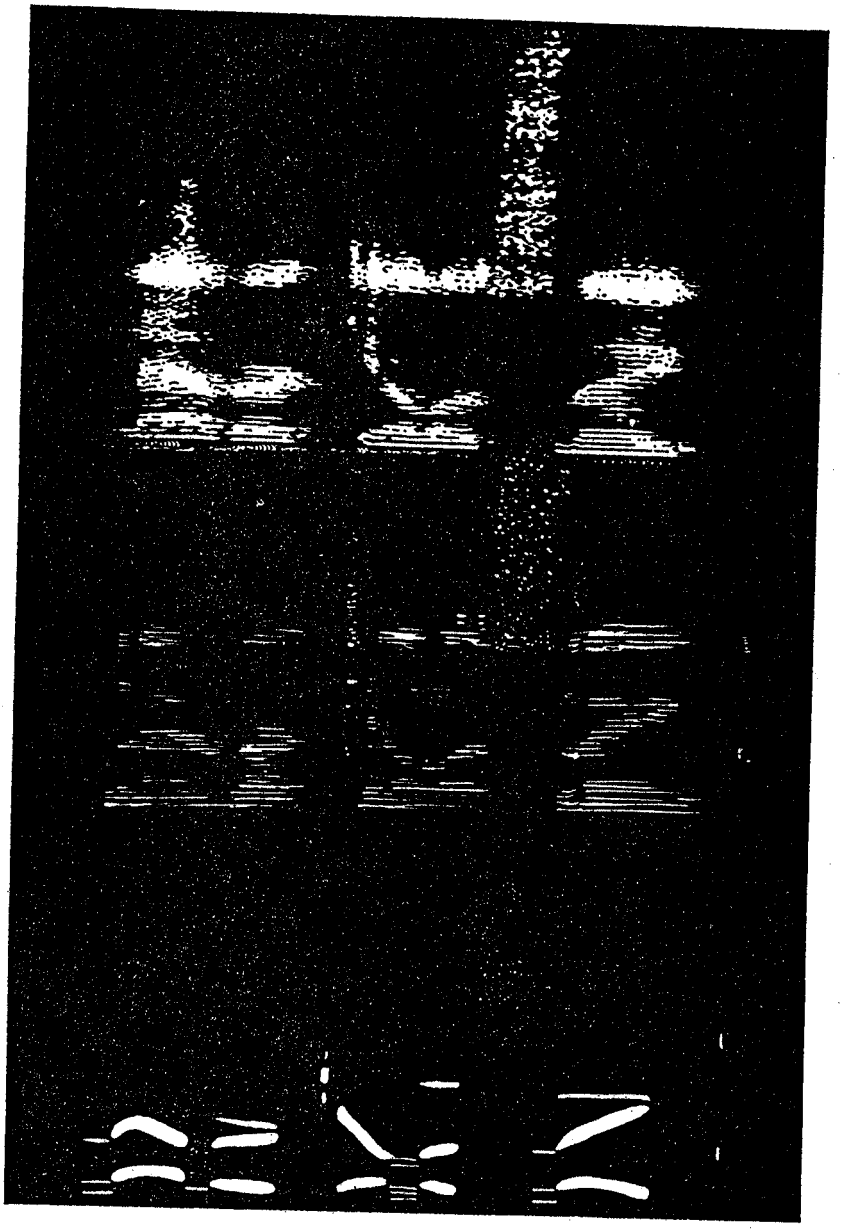

FIGURE 1

We have found, from studies of the sort we have just described, that it is possible to simplify and schematize the spectrographic pictures very considerably and still to have them just about as intelligible as the complete original spectrogram.

We turned next to a detailed study of individual speech sounds, attempting to strip them down to the simplest possible form even though this does result in sounds which are not as natural as those we could produce by copying spectrograms. With the vowels, we have worked out the frequency positions which give the correct vowel color when one limits the representation to two formants only. These cardinal vowels, some of which we will play for you now, were selected on the basis of rather extensive systematic variations of formant positions and relative intensities (see *Maître Phonetique,* December, 1951).

(Recordings of synthetic cardinal vowels [i], [e], [ɛ], [æ], [ɑ], [ɔ], [o], [u], were played.)

In experimenting with the stop consonants, we had, of course, to deal with dynamic aspects of speech. One characteristic of [p], [t], or [k] at the start of a syllable is the initial burst of noise. We have carried through one experiment in which the vowels were limited to two formants preceded by a burst of noise centered at each of several possible frequency positions.

A systematic test of each of seven vowels with each of twelve frequency positions for the burst of noise—that is, 84 syllables—served to locate the positions of the burst which most nearly resemble spoken syllables. The results show very clearly that the identification of the initial consonant

FIG. 2. Systematic investigation of the frequency of the burst of noise involved in the voiceless stop consonants, *p*, *t*, and *k*. TOP: An "outline" of the experiment. In part A are shown the frequency positions and extents of the twelve bursts; in part B, the frequency positions of the formants of the seven vowels; and in part C, one of the 84 test syllables. BOTTOM: Results shown as the distribution of *p*-, *t*-, or *k*- preferences. Thus, a high frequency burst is heard as *t* with all seven vowels; a lower burst is heard as *k* when it is just above the second formant of the vowel, or as *p* if it is elsewhere. (Reproduced by courtesy of the *American Journal of Psychology*)

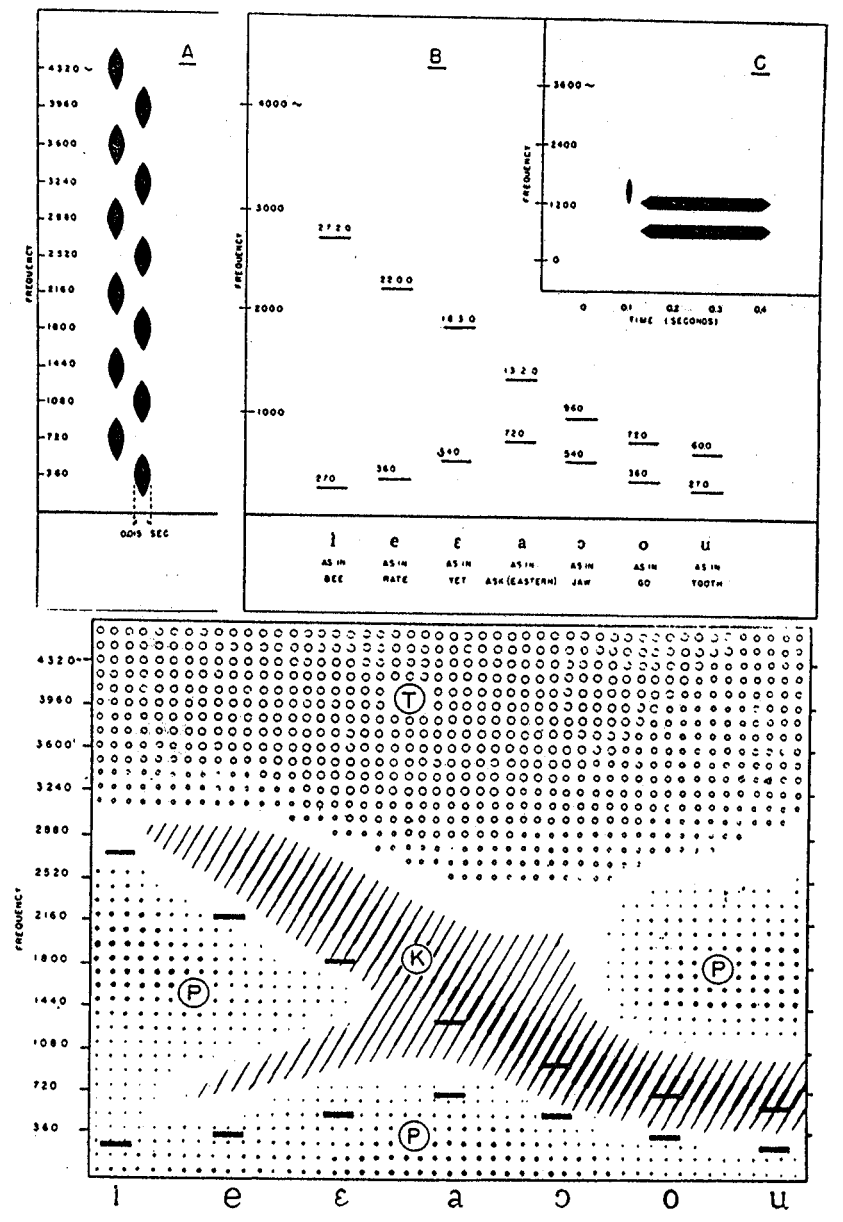
A
B
C
FREQUENCY
4320
3960
3600
3240
2880
2520
2160
1800
1440
1080
720
360
.005 SEC
4000
3000
2000
1000
2720
2200
1830
1320
960
720
600
270
360
540
720
360
270
l AS IN BEE
e AS IN RATE
ɛ AS IN YET
a AS IN ASK (EASTERN)
ɔ AS IN JAW
o AS IN GO
u AS IN TOOTH
3600
2400
1200
0
0
0.1
0.2
0.3
0.4
TIME (SECONDS)
T
K
P
P
P
l
e
ɛ
a
ɔ
o
u

FIGURE 2

does not depend *simply on the position of the burst, but on burst in relation to the vowel,* i.e., on the syllable.

(Diagrams taken from the complete account of this experiment, in the Oct., 1952, issue of *The American Journal of Psychology* were shown. See Fig. 2.)

But the initial burst of noise is not the only thing we see in the spectrograms for stop consonants. Usually the portion of the vowel immediately following the stop shows the formant in rapid transition. A second systematic experiment involved syllables without bursts preceding voicing, but with some eleven different degrees of initial transition, that is, from an extreme downward shift, to no transition, to an abrupt upward shift. This work is still in midcourse and we shall not attempt to go very far into the results which we have obtained thus far, except to say that the transitions alone provide adequate cues to the identification of stop consonants, and that the perception of the transitions is very much influenced by the following vowel.

(Some recordings were played illustrating the synthetic stop consonants differentiated by vowel transitions. See Fig. 3 for corresponding spectrograms.)

Ultimately, findings from research of this kind should enable us to synthesize speech without copying from spectrograms. We have not tried to go very far in this direction, but here is an example which may interest you.

(Hand painted spectrograms of the word *Alabama,* a) with American pronunciation, Southern accent, b) with French pronunciation, were shown (Fig. 4), and recordings of the corresponding sounds were played. These were painted in accordance with rules derived from our research, and not by reference to spectrograms of spoken sounds.)

FIG. 3. Vowel transitions as isolated cues for the perception of stops and resonants. The transitions shown are those which were most reliably identified by ear as the syllables indicated by the phonetic symbols.

FIGURE 3

FIGURE 4

Let us turn now to some aspects of this sort of research which seem to bear on the problem of teaching a second language. We shall point out some suggestions for teaching methods based on our experimental findings thus far. We trust that you will realize how very tentative these suggestions are.

1. *French nasals.* As you know, French has four nasal vowels, [ɛ̃], [œ̃], [ɔ̃], [ɑ̃]. Spectrographic analysis and synthetic reproduction of those four nasals shows that their articulatory position is not at all similar to that of the oral vowels whose symbols they share. For instance:

[ɛ̃] does not have the organs in the position of [ɛ]:
[ɛ̃] is less fronted and less open; when denasalized, it is not far from [œ].
[ɑ̃] does not have the articulatory position of [ɑ]:
[ɑ̃] is farther back and much less open; when denasalized, it is not far from [ɔ].

This was found by analysis of spectrograms in which denasalized nasals were compared with the oral vowels whose symbols the nasals share. By synthesis we were able to determine the acoustic features which had to be added to an oral vowel to make it sound nasal, and also the combination of oral and nasal features which gave the closest approximations to the four French Nasals.

The practical lesson here seems to be that the allegedly phonetic method of teaching the French nasals is not sound. The student who is taught to say [ɛ], then to nasalize it to [ɛ̃], does it correctly only in front of his teacher while he can have the sound repeated to him, so that he can make—unconsciously—the proper compensations of tongue and lip positions; when he practices nasalizing [ɛ̃] without the teacher to correct him, he is very likely to fixate an incorrect pronunciation.

FIG. 4. Two versions of the word "Alabama" painted by the rules which have emerged from systematic studies of groups of phonemes. TOP: With an American (Southern) accent. BOTTOM: With a French accent.

The correct way to teach French nasals would seem to be by direct imitation—preferably in words or sentences—independently of all oral vowels. Recordings or a good instructor are indicated.

2. *Movement vs. steady state.* Almost every day, in the laboratory, we meet new evidence of the importance of change, movement, as opposed to steady state, in the perception of speech. The steady state implied in the traditional descriptions of speech sounds can hardly be found in spectrograms, and in synthesizing speech we must constantly deal with changes—that is, frequency changes of the formants (in acoustical terms) corresponding to articulatory movements (in physiological terms).

In our research, we have found that at least three types of changes—change in extent, change in rate, change in direction—are important for identifying or discriminating speech sounds. All other conditions being equal, it is possible, for example, to distinguish, in synthetic sounds:

[ga] from [da] by extent of change of formant 2,
[an] from [aŋ] by extent of change of formant 2,
[am] from [al] by rate of change of formant 2,
[al] from [au] by rate of change of formant 1,
[ba] from [ga] by direction of change of formant 2,
[am] from [an] by direction of change of formant 2.

Another way to demonstrate the importance of these changes is, in playing back the spectrogram, to stop at several points in the course of a single phone. What we hear, then, is the steady-state sounds corresponding to those instants of time. Rarely do any of those correspond to the phone. Thus, stopping the spectrogram at different points in the [l] in *child* gives different vowel-like sounds ranging from about [o] to [u]; the short vowel [ɛ] of *leg* begins near [a] and ends near [I] without yielding a clear [ɛ] anywhere in between; and the different points of [b] in [ba] give a series of vowels ranging from about [u] to [a].

Another example of the importance of change in the perception of speech is offered by the voiced stops [b], [d], [g],

which can be made quite intelligible just by synthesizing the frequency changes—or so called transitions—to the contiguous vowel, the explosion itself being entirely omitted.

The practical lesson here is perhaps that the teaching of speech sounds as steady states may be largely useless. They should be taught *in movement* from and to the contiguous sounds, that is, in syllables, in words, and in connected speech.

3. *The Syllable.* The importance of change leads us to the third point, which concerns the syllable.

Two extensive experiments, one of them completed, the other in course, furnish strong indication that the irreducible acoustic stimulus of speech is not the phoneme but the syllable. (We do not mean that the phoneme is not the *linguistic* unit, but that its perception, in syllables of more than one phone, seems not to occur independently of the neighboring phone.)

In one experiment (briefly described earlier in this paper) we were looking for the preferred frequency of the burst of sound occurring in the production of initial [p], [t], [k], before each of the main vowels, [i], [e], [ɛ], [a], [ɔ], [o], [u]. Among other things, we found this interesting phenomenon: one of the bursts was heard as [k] before [a], but as [p] before [i] and before [u]. In other words the *same acoustic stimulus* was perceived in two ways depending on the neighboring stimulus.

In a second experiment (also mentioned earlier) we are investigating the effect on initial [b], [d], [g], of the *rate, extent,* and *direction,* of frequency changes (formant transitions) at the beginning of the vowel. We find that two transitions of same rate, extent, and direction, may be perceived differently depending on the vowel to which they are joined; for example, the rate, extent and direction of transition that is perceived as [g] before [e] is perceived as [d] before [ɔ].

In certain of these cases, therefore, it seems that the brain does not perceive the initial consonant of a syllable until the whole syllable has been heard, or in other words, that consonant and vowel are dealt with as a unit.

The practical lesson is again, perhaps, that isolated sounds should not be used in teaching, but only connected speech, or at least syllables, and preferably, of course, those that occur with high frequency in the language. A phoneme is known only after practice with the neighboring phonemes of the language.

4. *The role of articulatory movements in the perception of speech.* One explanation of the two phenomena just described lies in a motor theory of speech perception, that is, in the assumption that phonemes may not be perceived directly from the acoustic wave impinging upon the tympanum, but rather indirectly by reference to the proprioceptive stimuli which arise, or would arise, from the movements of articulation corresponding to those phonemes. The proprioceptive stimuli would be different for the two consonants compared ([k] and [p] in the first experiment, [g] and [d] in the second), because of the *articulatory* relationship of consonant and vowel, and therefore the perceptions would be different even though the acoustic stimuli were not different. Let us put it another way. In these two pairs of events, the perceptual event is more like the articulatory event. Therefore it is fair to assume that the articulatory event occurs as a link, or a basis for reference, between the acoustic and the perceptual ones.

The practical suggestion from this theory—and that is what a theory is for—is that while studying language we must not only listen but articulate—indeed, listen by articulating. The sounds produced by these articulations must be actually those of the second language. Not until correct habits of articulation are acquired, are we able to hear the second language correctly, let alone reproduce it.

(It might be added that the student should listen to his voice through a recording of it, so as to receive it from the outside, as he receives the teacher's voice from the outside. This simplifies for him the task of comparing his own pronunciation with that of the model, since it eliminates the modifications normally introduced in hearing his own voice through bone conduction.)

5. *Spectrographic displays as teaching aids.* One of the newer methods of teaching languages involves the use of the *spectrograph* and the *direct translator* as means of checking one's pronunciation. If this method is to be maximally effective we must know first which aspects of the spectrographic picture are important for the recognition of speech. As we pointed out earlier in this paper, it is difficult, and sometimes impossible, by simply examining the spectrogram, to determine the relation between what is seen on the spectrogram and what is heard. Our own research has been primarily concerned with the attempt to find these relations, by simplifying, and otherwise modifying, the spectrogram, and then determining the effects on the sound as heard. With this method, we are engaged in finding the essential acoustic cues to speech perception, the allowable range of stimulus variations, and, by exclusion, those acoustic components which are redundant linguistically.

We should hope that results from research of this kind will be applicable, not only to the use of the direct translator as a teaching device, but more generally to the problem of teaching a second language.

Finally, we wish to acknowledge the support of the Carnegie Corporation of New York and of the Department of Defense in connection with contract DA 49-170-sc-773.