Proc. 5th int. Congr. phon. Sci., Münster 1964, pp. 551–555
(S. Karger, Basel/New York 1965).

IBM San Jose Research Laboratory, San Jose, Calif.

Duration as an Alternate Synthesis-Parameter for Intensity and Vowel-Quality

By H. M. Truby

As I indicated at the Thursday morning plenary session, in my discussion of the paper of *Fant*, the specific purpose of a given phonetic evaluation *must* be made clear! It is rarely possible to determine with any validity the *acoustic* correlates of particular *physiological* activities, and as a result the phonetician plays many rôles in present day speech research.

My own long-range purpose as a committed phonetician is to identify as many particulars relating, in whatever way, to whichever aspect of speech, as circumstance permits. Therefore I am interested in anatomic details, in articulatory details, in acoustic details, and in the ways in which these details relate to the production, transmission, and perception of speech.

This afternoon I come to you as a bearer of acoustic tidings from a machine. Please keep in mind the restrictions, limitations, and objectives that such a mission indicates.

The phonetician studies the sounds of speech – but rarely to the exclusion of context. The so-called "isolated vowel" appears in a context of silence and with a specific reference to other linguistic planes. The phonetician who deals with *synthesized* speech, operates with examples whose identification is absolutely dependent upon the effectiveness of the *caricature* nature of the presented information.

Allow me the pleasure of an analogy: The following little sketch is, in a basic way, a caricature

of a tulip... of a man... of a woman...

A synthesized version "of a vowel" is a *caricature*, not only of "the vowel", but as well of a particular phonetic context, semantic context, grammatical context.

I ask your attention to a few such examples of speech caricature: It is not difficult to demonstrate the intelligibility of synthesized-speech-using-familiar-context-alignments, e.g., /[ˌgutn̩ 'tak vi 'getɪs ˌinɛn]?/, /[ˌbɔ̃ 'žuR 'kɔ̃mɔ̃n, ta le vu]?/, /[ˌd*a*s ˌdoɪ 'k*a*po sa v*a*-ɪš]/, /['hɛlo ˌhaʊ '*a*r ju]?/. Also readily demonstrable is the contribution of particular parameters to this intelligibility by a now familiar technique especially adaptable to the IBM Terminal Analog Speech Synthesizer: the first sample comprises only the acoustic information contributed by a monotoned F_0 with controlled timing ((1)); for the next sample F_1 information is added: ((2)); next we hear the sample with F_2 information added: ((3)); then with the addition of F_3 information: ((4)); then with F_0 controlled so as to simulate one of many possible intonation patterns: ((5)).

Now, it is, clearly, optimistic to expect consistently positive identification, by listeners naïve to the voicelike characteristics of a particular speech-synthesizer, of isolated simulated "words", but by the process of elimination and/or with a little practice, identifications do approach consistency. And then, too, there is the helpful – and at the same time frustrating – *perceptual* phenomenon of "knowing what the word *is*", thereby providing cues which serve as well as – or even better than – what might be termed "valid acoustic cues". For example, here are three synthesized words from our inventory which have, under ideal and even less-than-ideal listening conditions, enjoyed consistent identifications:

/[wɪnd]/ × 3   /[did]/ × 3   /[dɪd]/ × 3

It is a commonplace of phonetics that in the process identified by the term "vowel gradation", the reduced vowel *loses* its "original" vowel color and tends toward a "more neutral" vowel. Traditional observations in this regard posit four specific "neutral vowel regions" for unstressed vowels: in the "front-vowel region" the tendency is toward [ɪ], for example, /ri-/ is perceived as [rɪ-] in [rɪ'pit] (*repeat*); in the so-called "back-vowel region" the tendency is toward [ʊ], e.g. /tu-/ is perceived as [tʊ-] in [tʊ'de] (*today*); and the so-called "central-vowel region" is either retroflexed schwa [ə̣] or schwa [ə] depending upon the circumstances, e.g., note how /pɜ̣-/ is manifested in [pə̣'siv] (*perceive*), and how /ʌ/ becomes schwa in [ə'lon] (*alone*).

Schwa is, of course, an extremely popular linguistic notion, but the phonetician should not be so willing to accept this phonetically-broad generalization. It is certainly not new to phoneticians that, in a manner of speaking, schwa manifests itself in a wide variety of differing phonetic forms, and my own proposed title for this phenomenon is "And a Little Schwa Shall Lead Them".

In any case, certain gradations *seem* obvious, and I single out one of these for this report:

The identity of a reduced vowel is contingent upon the degree of stress of the carrier syllable as well as upon the degree of stress of the contrastive or tonic syllable. This degree of stress – or "unstress" – of the *carrier* syllable influences the phonetic character – the spectral shape in time – of all members *of* the syllable, and similarly for the members of the *tonic* syllable.

Let us take advantage of the fact that certain controls *can* be effected in speech synthesis which can*not* be effected in natural speech due to interrelated, involuntary, compensatory adjustments throughout the phonetic environment as a whole. My objective is to manufacture the following caricatures: Begin with /['kæn'did]/ (*Candide*, a play by G. B. Shaw), and systematically reduce the duration of the /-i-/-vowel until /['kændid]/ (*candied*) is heard, and eventually /['kændɪd]/ (*candid*). (Since it is clear that the minimal acoustic cues provided by speech-synthesis relate only by perceptual similarity to speech, the transcription feature /[--]/ presented in my *Acoustico-Cineradiographic Analysis Considerations*, Acta Radiologica Supplementum 182, Stockholm, 1959, is employed here).

Synthetic "utterances" are contrived with our present system by assembling sequences of what I have termed *diphones*. For example: kæ + æn + nd + di + id. Essentially, *diphone nuclei*, representing transitions, are stoı ed in the computer, and absolutely steady-state formants are extrapolated by the computer from the relevant end-point of the particular nucleus. The systematic exploitation of each of all possible control parameters is then only a matter of computer programming. The evaluation is up to the phonetician:
For example,

[kæ–	–æn–	–nd–	–di–	–id]		
180	140	70	50	80	=	kændid
				85	=	kændi·d
				90	=	kændi:d
				95	=	kændi:·d
	145				=	kæn·di:·d
				90	=	kæn·di:d
				85	=	kæn·di·d
				80	=	kæn·did
	150				=	kæn:did
				85	=	kæn:di·d
				etc.		
	etc.				with	kæn:·
etc.					with	kæ·
					with	kæ:
					with	kæ:·

((etc., etc.))

The figures cite diphone durations in milliseconds.

In conclusion may I offer the observation that, being phoneticians, we welcome the opportunity to "hear for ourselves" exempla ordinarily "listener-tested" and presented as statistics. As Dr. *Cooper* indicated, we can now call on the computer to provide us not only with the acoustic samples for our listening pleasure but with correlated print-outs such as these six-foot sheets of data I enfold before you.

Author's address: Dr. H. M. Truby, Communication Research Institute, *Miami, Florida* (USA).

Discussion

Isačenko (Berlin): Sie haben simulierte Rede als «Karikatur» bezeichnet. Wenn man ein Symbol, welches nur die relevanten Züge eines Objekts darstellt, als Karikatur bezeichnet, so mögen Sie recht haben. Ist aber der Klavierauszug eine «Karikatur» einer Oper oder einer Symphonie? Ist die Wiedergabe der Rede durch ein schlechtes Telephon die «Karikatur» der Rede? Ich glaube, daß die Darstellung der *relevanten* Züge eines Objekts dieses Objekt erschöpfend darstellt.

J. E. Damman (New York): Do you think the results achieved in this case reflect at all the constraints imposed by having a choice of only two words in English: "candied" and "candid"? That is, do you feel the seeming phonetic shift will carry over when a clear cut but restricted choice of this sort is not present?

Answer Truby to Mr. *Isačenko's* reluctance to accept synthesis speech within my definition of *caricature*, I pointed out that by selecting minimal spectral components and emphazing those particular components which insure phonemic identification of a particular phonetic segment, I am, certainly, producing a caricature of some Hochsprache segment.

To *Dammann:* The demonstrated "phonetic shift" will operate whenever the conditions of contrast indicated are operative.