Proceedings of The Institute of Acoustics

RAPID VERSUS SLOW SPECTRAL CHANGE: IMPLICATIONS FOR DYNAMIC AUDITORY PROCESSING OF SPEECH

Anthony Bladon

Phonetics Laboratory, University of Oxford,
41 Wellington Square, Oxford, OX1 2JF

Early speculations about the nature of dynamic auditory processing of speech began with the simplest possible step from existing knowledge about steady-state processing. One idea floated was that the perceptual system would accumulate a running comparison, over time, between incoming speech and stored reference patterns. Thus the spectrum samples would be collected at frequent intervals and integrated over a longish interval, as much as a diphone long. This scenario was put forward tentatively by Klatt (1979), and was explored at some length by the Leningrad group (Chistovich et al., 1982). However, despite its respectable pedigree, the idea is quite obviously inadequate. It will not do to suppose that perceptual sampling runs uniformly along over time. Speech listening entropy varies greatly at different instants. Cues to a lateral consonant, or a stop consonant, lie less in their steady-state than in the transitions into and out of their steady states, although the transitions are briefer. Clearly, the brain weights more strongly the period of spectral change.

A slightly more sophisticated view of the perceptual treatment of spectral change was put forward by me in the Fallside and Woods volume, still awaited (Bladon, forthc.). It was suggested that spectral rate-of-change should itself be used as the weighting factor by which the auditory sampler would adjust its activity. Thus at times of rapid spectral change, such as in a trill, a nasal release, or in a plosive burst, the spectrum would be most densely sampled; and during intervals of spectral stability, some low sampling rate would apply.

In this paper, I survey evidence relevant to this view of processing of speech as weighted by spectral rate-of-change. I now find this view only minimally satisfactory. The evidence indicates that we need in addition to provide differently for spectral change consisting mainly of energy offsets, and for that consisting mainly of energy onsets. I also identify a need for a second type of role for spectral change, one where the interval of change is itself not inspected for its intrinsic spectral content, but rather the spectral change acts as a dynamic flag and as a temporal pointer to adjacent spectrally interesting regions. (I will expand on these notions.) It seems that this dichotomy of role for spectral change in speech can be broadly equated with rapid transitions versus slow ones. It will be suggested that the rapid transitions contain information which is perceptually important, to an extent which the slow transitions do not.

RAPID VERSUS SLOW SPECTRAL CHANGE: IMPLICATIONS FOR DYNAMIC AUDITORY PROCESSING OF SPEECH

Let us therefore turn first to a case which seems reasonably well fitted by the moderately crude (Bladon, forthc.) conceptualization of the role of spectral change. Stop consonants (or more strictly, stop-vowel sequences) do seem to fit the idea that perceptual information is accumulated in proportion to the instantaneous rate-of-change. Of course, this is an area where there has long been a division of opinion, into what might be termed the burst camp and the formant transition camp. I see every sign of a reconciliation, however. From the work of Fischer-Jorgensen (1972) and Kuhn (1979) who explored the differential role of these burst and transition components in stops, it is possible to conclude that both components contribute importantly to the place percept, although each cue does not contribute uniformly to every place. A further token of compromise is that Blumstein's more recent pronouncements reflect some shift from her camp-bound position: while the (immortal?) burst templates are still a satisfactory way - probably still an invariant way - of discriminating among places, she concedes that the spectral shape of the stop burst is not its primary perceptual cue (1982: 49).

So what does the auditory system make of this multiplicity of cues (for there are of course others too)? There is good evidence that it integrates them in some fashion. This has been argued experimentally by Sawusch and Pisoni (1974), and modelled theoretically by Oden (1978), and it pervades much of the recent Repp output. However, the integration window will have to be rather variable. This was demonstrated by Kewley-Port (1982) who found that auditory-spectral templates à la Bladon and Lindblom (1981) for a velar stop burst had to persist over a longer time-window than the templates for the other stops. Searle et al. (1979), also using quasi-auditory spectra, amplified the point: based on their identification results, the stimuli they construct have a 90 ms burst for /kV/ but only 20 ms for /tV/. From this rather persuasive evidence, and assuming that the perceptual system needs to be able to identify a /kV/ and a /tV/ roughly equally well, I conclude that the auditory sampling during the /tV/ burst should be roughly as dense as that over the longer /kV/ burst, and to achieve this, a weighting by rate-of-change seems appropriate.

For formant transitions, into and out of a stop or a nasal, a similar argument can be made. Such rapid transitions are apparently processed spectrally in a cumulative manner. The auditory image they create can probably undergo an enhancement of frequency information due to lateral supression (Lacerda and Moreira, 1982). The temporal profiles of auditory nerve discharge follow rapid formant transitions fairly faithfully (Sachs et al., 1982). Their perceptual cueing value is not in doubt. And so, I would conclude up to here, a reasonable approximation to the dynamic processing of these transitional stretches of speech is one where the spectrum is sampled, during its change, at a rate proportional to its rate-of-change.

However, in an important respect the above presentation is too naive. It would predict equivalent auditory importance for 'mirror image' spectral changes. An example of a 'mirror image' would be the highly similar way in which the physical spectrum changes at (a) release of a fricative onto a vowel and at (b) closure of a fricative after a vowel. However, one of the emerging facts about temporal auditory analysis is what may be termed an onset/offset asymmetry. There is both psychoacoustic (Tyler et al., 1982) and physiological (Delgutte and Kiang, 1984) evidence for this. For example, because of such properties as short-term adaptation in the auditory nerve, onsets (higher density of spike discharge) are considerably more strongly represented than offsets. A good illustration can be given of how this asymmetry may surface in speech patterns of languages. Phoneticians have widely attested (Ruhlen, 1978) that the spread of nasality from a nasal onto an adjacent vowel is much more pervasive in the sequence VN ('anticipatory nasalization') than in the opposite sequence NV. Now the 'boundary' in the sequence VN consists largely of weak spectral energy offsets, and so is particularly vulnerable to the adaptation phenomenon, causing auditory temporal smear. Thus the spread of nasality is facilitated. (The opposite sequence NV shows weak spectral energy change also, but at least these are onsets!) What this example teaches us, then, is that the modelling of spectral change must be sensitive to more than just the rate-of-change, giving extra points, if you wish, to a positive-going sign.

Now I come to my second dissatisfaction with the hypothesis that spectral change in speech is sampled according to its rate-of-change. This is in the case of speech events in which the transition is relatively slow. As my archetype example, I take diphthongs. Now it may be reasonable to say of stops, as was done above, that important information resides in the spectral change interval; but diphthongs, I suggest, are identified not by sampling their changing spectrum at all.

Readers familiar with the diphthong literature, especially the Gay study (1970), may regard this last statement as heresy. For according to Gay, the American English diphthongs /ɔi, ai, au/ "are characterized primarily by an invariant rate of formant frequency change." This view would say, then, that the dynamic auditory analyser must sample the diphthong spectral change densely, to determine its rate-of-change (and hence the diphthong's identity). But this view is controversial. Less well known data by Chistovich et al. (1982) and by Holmgren (1979) are largely supportive of another position, namely that listeners identify these formant frequency dynamics not by rate-of-change (formant gradient), but by the trajectory endpoints alone.

I recently carried out a further set of experiments to try to illuminate this conflict. Summarizing here, diphthongs whose end was cut back were identified not in terms of the rate-of-change of the trajectory remaining, but in terms of the endpoint actually achieved. Also, and consistent with that finding, diphthongs which

had their transition totally deleted, thus consisting of a foreshortened step-like vowel sequence, were identified trivially easily. In many cases, to my surprise and despite the sudden shift of quality in the stimuli, subjects even reported there was no detectible manipulation of the stimulus at all. By contrast, to try to identify a diphthong from its transition alone led to a high proportion of errors.

My conclusion from these experiments was that diphthong information is carried principally by their endpoints (even though there may be very little steady-state). One thing the ear is not doing, during the diphthong transition, is determining the rate-of-change.

This is not to say, however, that the spectral change in a diphthong is not perceptually significant. It is important perceptually to know that change is taking place: we may presume that the change signifies "diphthongization is going on" (but not, "this is a candidate for diphthong x"). Several pieces of evidence contribute to the positive part of this assumption. We know for example that there are auditory mechanisms specifically tuned to respond to change (Møller, 1982). Consequently it is not surprising that confusions between a steady-state vowel and a diphthong are rare, or that identification scores for vowels with diphthongization are higher than those for steady-state vowels (Assmann et al., 1982).

In the last paragraph I have identified a role for spectral change, during the slower diphthong transitions, which might be thought of as a weighting flag. Thus, the very fact that there is change, even if it be only a slight diphthongization as in [iI], is auditorily very salient. While this slower spectral change flags for extra perceptual weight, then, it does not determine spectral distances in any interesting way. Instead, the role of the spectral change also involves acting as a pointer to temporaly adjacent regions of the signal, which are to be inspected for their spectral shape. In this case, these are the diphthong trajectory endpoints, the spectral sampling of which is, as I outlined, crucial to the diphthong's identification.

Some speculation that, in the interpretation of physiological data from the auditory nerve, a temporal pointer role may be relevant, is to be found in Delgutte and Kian (1984).

Other speech events might be brought briefly into the arena. Perceptual phonetic evidence enables us to begin to align these other events with one or other type of role which their spectral change plays. Like the stops, for example, are nasals and probably affricates. Like the diphthongs, on the other hand, appear to be the glides [w] [j] etc., which according to the latest evidence (Shinn and Blumstein, 1984), are distinguished from stops by amplitude difference and not by the glides' distinctive rate-of-frequency-change, which the perceptual system does not seem interested in computing. Laterals are a matter of some speculation. They do offer an interesting case where part of the steady-state spectral content in itself, namely the antiformant notches, is apparently

of no interest in a static form: the auditory filter appears to smoothe such notches over (Carlill, personal communication, see also Lacerda (1980) on the same effect in fricatives). But the antiformants may come into their own perceptually when they form a spectral discontinuity with adjacent vowel formants. Acting as a temporal pointer to the contrast formed by adjacent parts of the signal, then, the LV spectral change may perhaps be of the second type I have identified.

In summary, I have argued from a variety of kinds of evidence that the dynamic analysis of auditory spectral change in speech may be differential. Spectral change may be (a) intrinsically salient, in proportion to its instantaneous rate-of-change, (b) salient only as a flag, and simultaneously as a temporal pointer, and (c) in either case, asymmetric as between its positive (= onset) salience and negative (= offset) salience. There is a suggestion that the more rapid spectral changes in speech (at stop or stop-like boundaries) are interpreted predominantly in fashion (a), and the slower spectral changes (diphthongs, glides, perhaps lateral boundaries) function more obviously as in (b).

REFERENCES

Assmann P.F., Nearey T.M. and Hogan J.T. (1982). Vowel identification: Orthographic, perceptual and acoustic aspects. J.A.S.A. 71, 975-989.

Bladon R.A.W. (forthc.) Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: a thread. In Fallside F. and Woods W. (eds), Computer Speech Processing. Cambridge: CUP, Chapter 2.

Bladon R.A.W. and Lindblom B. (1981). Modeling the judgement of vowel quality differences. J.A.S.A. 69, 1414-1422.

Blumstein S.E. et al. (1982). J.A.S.A. 72, p. 49.

Chistovich L.A., Lublinskaya V.V., Malinnikova T.G., Ogorodnikova E.A., Stoljarova E.I., Zhukov S.J. (1982). Temporal processing of peripheral auditory patterns of speech. In Carlson R. and Granström B. (eds.), The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier Biomedical, pp. 165-180.

Delgutte B. and Kiang N.Y.S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. J.A.S.A. 75, 897-907.

Fischer-Jorgensen E. (1972). Perceptual studies of Danish stop consonants. Ann. Rep. Inst. Phonet. Univ. Copenhagen 6, 75-176.

Gay T. (1970). A perceptual study of American English diphthongs. Lang & Speech.

Holmgren K. (1979). Formant frequency target versus rate of change in vowel identification. Phonet. Exper. Res. Inst. Ling. Univ. Stockholm, 1, 83-91.

RAPID VERSUS SLOW SPECTRAL CHANGE: IMPLICATIONS FOR DYNAMIC AUDITORY PROCESSING OF SPEECH

Kewley-Port D. (1982). Measurements of formant transitions in naturally produced stop consonant-vowel syllables. J.A.S.A. 72, 379-389.

Klatt D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. J. Phonet. 7, 279-312.

Kuhn G.M. (1979). Stop consonant place perception with single formant stimuli: evidence for the role of the front cavity resonance. J.A.S.A. 65, 774-788.

Lacerda F. (1980). Application d'un modèle auditif à l'étude des confusions des fricatives non-voisées. Univ. des Sciences Humaines de Strasbourg. XIème Journées d'Etude sur la Parole, 239-248.

Lacerda F. and Moreira H.O. (1982). How does the peripheral auditory system represent formant transitions? A psychophysical approach. In R. Carlson and B. Granström (eds.), The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier Biomedical , 89-94.

Møller A.R. (1982). Neurophysiological basis for perception of complex sounds. In Carlson R. and Granström N. (eds.), The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier Biomedical , pp. 43-60.

Oden G.C. (1978). Integration of place and voicing information ... J. Phonet. 6. 83-93.

Ruhlen M. (1978). Nasal vowels. In Greenberg, J.H. (ed). Universals of Human Language, Vol. 2: Phonology. Stanford: Stanford University Press , 203-241.

Sachs M.B., Young E.D. and Miller M.I. (1982). Encoding of speech features in the auditory nerve. In R. Carlson and B. Granström (eds.), The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier Biomedical , 115-130.

Sawusch J.R. and Pisoni D.B. (1974). On the identification of place and voicing features in synthetic stop consonants. J. Phonet. 2, 181-194.

Searle C.L., Jacobson J.Z. and Rayment S.G. (1979). Stop consonant discrimination based on human audition. J.A.S.A. 65, 799-809.

Shinn P. and Blumstein S.E. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. J.A.S.A. 75, 1243-1251.

Tyler R.S., Summerfield Q., Wood E.J. and Fernandes M.A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. J.A.S.A. 72, 740-752.