LOCATION AND RECOGNITION OF PLOSIVE CONSONANTS IN CONTINUOUS SPEECH

MARY O'KANE

School of Information Sciences and Engineering
Canberra College of Advanced Education
Belconnen 2616, Canberra, Australia

ABSTRACT

Algorithms to locate and classify plosive consonants in continuous speech have been incorporated and tested in the FOPHO continuous speech recognition system [1]. These algorithms are based on a detailed study of the speech of ten Australian English speakers (five male, five female) who participated in a word game in which each speaker produced continuous speech versions of all possible VCV combinations where the vowels were either the high front vowel /i/ or the low back vowel /ɔ/ and where the consonant was one of the six plosive consonants of Australian English /p,b,t,d,k,g/. The study showed that a complete plosive classification algorithm must be context-dependent because it was found that generally all speakers produced the plosives in a heavily coarticulated manner with systematically varying coarticulation phenomena being observed in formant transitions, timing effects, bursts and pseudo-loci as one progresses from bilabial through alveolar to velar plosives. Another important factor that has to be taken into account in the recognition algorithm is the speaker's sex.

INTRODUCTION

An algorithm for the classification of plosive consonants in continuous speech was developed from a study of ten speakers, five male and five female, producing plosives in set contexts. This algorithm was tested and then generalised and incorporated into the FOPHO speech recognition system. In this paper the development of this algorithm is described and some results of using it in the recognition system are given.

FEATURES OF PLOSIVE CONSONANTS

When developing a recognition algorithm for the plosive consonants one has a large amount of literature of which to draw in order to determine what features of the plosive consonants are likely to be important for distinguishing between the plosives produced at different places of articulation. This literature can for convenience be grouped in three categories: perception experiments using synthetic speech, perception and production experiments using real speech, and classification studies. Synthetic speech experiments have highlighted the fundamental characteristics of plosives while the real speech experiments have tended to clarify the interaction of these characteristics and the variability of their realisations in real speech. Automatic recognition studies give some guide as to which features can be located automatically with reasonable efficiency and robustness. Here we give only a very cursory guide to these three types of literature.

In synthetic speech experiments in the early 1950's Cooper, Delattre, Liberman, Borst, and Gerstman [2] found that certain stop burst spectra were characteristic of the three stop types. In later experiments Liberman, Delattre and Cooper, [3], found that the formant transitions from the plosive consonant to the following vowel produced a successful synthetic voiced plosive. The first formant transition appeared to contribute to the voicing of the stop while the second formant transition provided a basis for distinguishing between the stop types. Further experiments led to development of the now famous 'locus theory' which postulates that the second formant transitions should point to a frequency locus no matter what the following vowel is. Delattre, Liberman, and Cooper [4] found that this was particularly characteristic of /d/, not quite so reliable for /b/, while for /k/ there were two loci, a high one if the following vowel was a front vowel and a low one if the following vowel was a back vowel. Hoffman [5] further refined much of the previous plosive research on stop consonants to see how the two main types of cues for plosive conconants, burst and formant transitions, interacted. He concluded that all the cues are perceptually independent of the other cues present and that for some stops, notably /b/ and /g/ the burst provided a weak cue and the transitions a strong cue while for /d/ the burst was a strong cue and the transitions were a weak cue. Later research has demonstrated the significance of these conclusions.

In real speech experiments, Halle, Hughes and Radley, [6] considered plosives occurring not only in conjunction with vowels but also in consonant clusters and at the beginnings and ends of words. They concluded that a complex array of cues was needed to characterise the plosives and that the locus theory was somewhat inadequate for this task. In this they were supported by Öhman [7] who studied spectra of VCV coarticulations for Swedish, using all possible combinations in which the consonant was a plosive. He deduced that each VCV coarticulation

was a 'basic dipthongal gesture with an independent stop consonant gesture superimposed on its transitional portion'. The relative importance of burst and transitions in signalling plosive consonants has been heavily debated. However there is evidence (see [8] for example) that different speakers signal various plosives in different ways. Despite this Stevens and Blumstein [9] showed that stimuli as short as 10-20 msec sampled from the onset of consonant-vowel syllables can be reliably classified according to place of articulation using gross spectral shape wave 85% of the time. Studies such as that by Lisker and Abranson [10] have shown that timing phenomena such as VOT are also crucial to the correct production of plosives.

There have been several published descriptions of algorithms for the automatic discrimination of plosive consonants in various languages (e.g. [6], [11], [12], [13], [14], [15], [16], [17]). Considering this group of algorithms as a class, the most favoured aspect of the plosive for plosive discrimination is the burst which is often analysed according to some frequency-band scheme. Measurements of transitions and timing tend to be used as secondary recognition cues.

PLOSIVES IN CONTINUOUS SPEECH

Many of the published algorithms for the recognition of plosive consonants were developed from studies of plosives produced in citation syllables. From studying the literature described in the previous section we concluded that citation-form syllables were unlikely to be fully representative of the range of plosive production phenomena that speakers might use in continuous speech. Therefore, in order to develop an algorithm for the recognition of plosive consonants in Australian English continuous speech we designed an experiment in which both male and female speakers produced all the plosives occurring in English in a range of VCV coarticulatory settings with the added complication of junctural effects occurring within the VCV triplet.

This experiment was conducted as follows:
Lists of two-word sequences were prepared. Each of these two-word sequences was one of two forms:

(1) The first word ended in a VC combination and the second word began with a V; where the vowels could be either the high front vowel /i/ or the low back vowel /ɔ/ and the consonant was one of the six plosives e.g. 'heat ought', 'morgue awful';
(2) The first word ended with V and the second word began with a CV combination where the vowels could be either /i/ or /ɔ/ and the consonant was one of the six plosives e.g. 'he taught', 'more gory'.

Thus with the two vowels and six plosive consonants and two juncture positions there were a total of forty-eight two-word combinations. Five male and five female speakers who all spoke standard educated Australian were each presented with the list of two-word sequences and instructed not to study the list but to immediately begin saying sentences containing the word sequences. It was impressed on the subjects that the sentences they produced were to be spoken at a conversational speed and that the semantic content of the sentences was of no particular importance. This was to keep the subject speaking at as conversational a rate as possible. This aim was largely achieved. With this experimental paradigm the phonetic and junctural contexts were controlled but the speech used was reasonably representative of continuous speech.

The sentences containing the two-word sequences were recorded and then the required VCV tokens were excised and digitised (with a 10 kHz sampling rate for male voices and 16 kHz for female voices). These tokens were then analysed using an autocorrelation-based linear prediction technique. Timing, formant transition and burst phenomena were all investigated. The results of these investigations are briefly described below.

TIMING PHENOMENA

The particular timing parameter measured, chosen because it was easily amenable to automatic measurement, was the interval which began at the point in time corresponding to the minimum gradient point of the waveform rms energy curve in the region in which the rms energy decreases from its value for the steady state of the vowel to the closure for the stop consonant, and ended at the point in time corresponding to the point of maximum slope of the waveform energy curve in the region in which the rms energy curve increases after the plosive burst to its (much higher) value during the steady state of the vowel following the consonant. It was found that this measurement reflected a complex interaction of junctural, voicing, place of articulation and speaker differences. Some of these effects are illustrated in figure 1. This interaction of effects meant that this timing parameter is of little use as a primary recognition determiner although it can be used as a check on extreme cases.

FORMANT TRANSITION PHENOMENA

A detailed description of the results for formant transitions has been given elsewhere [18]. In summary, it was found that the second formant transitions did indeed display strong locus effects with a low locus range for labial plosives, an intermediate locus range for alveolar plosives and two locus ranges for velar plosives - a high locus range if the vowel receding the consonant was the high front vowel /i/ and a low locus range if the vowel preceding the locus range was the low back vowel /ɔ/. Indeed it was found that the position of the locus was primarily determined by the vowel preceding the consonant with the vowel following the consonant having a modifying effect on this locus. Within sex groupings inter-speaker differences of locus positions were slight while the differences between male and female locus

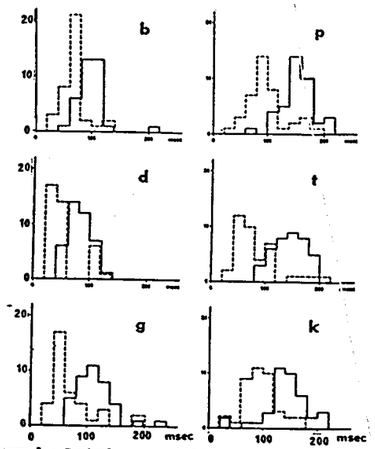

Figure 1 : Typical examples of bilabial, alveolar and velar bursts for male speakers in a/ɔ-ɔ/ context.

positions are reflections of the differences between vowel second formant positions for male and female voices.

Transition timing effects were also investigated. Transition lengths were found to be somewhat dependent on the place of articulation of the consonant. In particular it is very noticeable that transitions from vowels to the velar plosives (/k/, /g/) and from velar plosives to vowels are on average much shorter than transitions to and from labial and alveolar plosives.

BURST PHENOMENA

It was found that the burst spectral shapes fell into the three categories. Labial bursts were diffuse and of low level energy. Alveolar bursts were generally not so diffuse and were higher level energy bursts than the labial bursts. Alveolar bursts are most prominent in the 2.5-4 kHz region while labial bursts tend to be in the 1-3 kHz region. Velar bursts characteristically display two narrow-bandwidth peaks. The more prominent of these occurs in the 0.7-2.8 kHz region, and the (generally) smaller one occurs in the range 3.5-5 kHz. The exact position of these peaks is dependent on the nature of the surrounding vowels. Typical examples of the three burst shapes are given in figure 2.

With regard to the effects of coarticulation and the burst spectrum the most noticeable effect is that coarticulation effects are more evident and more consistent in velar than in alveolar plosives. And coarticulation effects are more evident and more consistent in a alveolar plosives than in bilabial plosives. Another effect is that different speakers can manifest coarticulation effects in bursts in a variety of ways and to a variety of degrees. However coarticulation effects in velar bursts are quite spectacular for all speakers. Such effects are most strongly indicated by the position of the lower in frequency (and generally higher in amplitude) of the two peaks of the typical velar spectrum. An example of this phenomenon is displayed in figure 3. Male/female differences in bursts again tend to reflect male/female differences in vowel formants. It should be noted that 7% of voiced plosive consonants were produced without any discernable burst. Certain speakers are more prone to this mode of production than others. Generally however, these speakers produce very clear formant transitions.

A PLOSIVE RECOGNITION ALGORITHM

A algorithm for automatically classifying the plosive consonants was developed from this study. In particular it was developed from the results for eight speakers and tested on the remaining two. This algorithm is described in detail in [19]. The rules were speaker-independent within each sex grouping and produced a fuzzy estimate of the likelihood that any unknown plosive was produced in any of the three possible places of articulation. The rules primarily involved three fundamental measurements - measurements of bursts, measurements of formant transition endpoints and measurements of formant transition slopes. The rules are constructed such that a variety of speaker variation effects (such as burst non-production) are allowed for. Using these rules it was found that 92% of the plosive consonants tested were correctly and uniquely classified with membership greater than or equal to 0.5. The velar consonants were the most successfully recognised class of sounds with an average of 95% correct recognition. The labial consonants were recognised on average 90% of the time and the alveolar consonants were correctly recognised 91% of the time. In about 6% of all cases consonants were classified correctly but also received a simultaneously high rating in an incorrect category.

GENERALISING THE RULES

The plosive classification rules described above were incorporated in the FOPHO recognition system along with a plosive location algorithm which depended primarily on burst location and to a lesser extent on energy contour and timing effects. It was found that the overall recognition of plosives was poorer in the more general situation and that many of the refinements in the classification rules (such as those allowing for burst non-production) were rarely not invoked as many plosives productions that might have been classified by these rules were not located. Thus the usefulness of studies such as the one described here is limited unless plosive location algorithms are good.

Also as only about 25% of all plosive productions in continuous speech occur in VCV contexts the sophisticated coarticulation phenomena noted for these situations is only of limited immediate usefulness. Nevertheless they have proved a useful guide for other contexts and results for

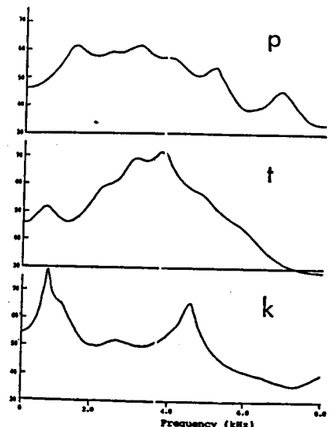

Figure 2 : Histograms showing the distribution of the timing parameter. The dotted lines represent the /VC v/ cases and full lines represent the /V CV/ cases.

VCC contexts (where the plosive is the consonant after the vowel) tend to confirm that the vowel before the consonant is the primary determiner of transition coarticulation phenomena.

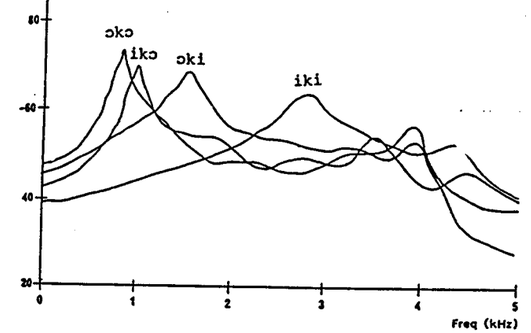

Figure 3 : Velar bursts showing coarticulation effects for a male speaker.

REFERENCES

[1] M. O'Kane, 'The FOPHO Speech Recognition Project', Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 1983, pp.630-632.

[2] F.S., Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst & L.J. Gerstman, 'Some experiments in the perception of synthetic speech sounds', J. Acoust. Soc. Amer., 24, 1952, pp.597-606.

[3] A.M. Liberman, P.C. Delattre & F.S. Cooper, 'The role of selected variables in the perception of the unvoiced stop consonants', American Journal of Psychology, 65, 1952, pp.497-516.

[4] P.C. Delattre, A.M. Liberman & F.S. Cooper, 'Acoustic loci and transitional cues for consonants', J. Acoust. Soc. Amer., 27, 1955, pp.769-773.

[5] H.S. Hoffman, 'Study of some cues in the perception of voiced stop consonants', J. Acoust. Soc. Amer., 30, 1958, pp.1035-1041.

[6] M. Halle, G.W. Hughes & J.P.A. Radley, 'Acoustic properties of stop consonants', J. Acoust. Soc. Amer., 29, 1, 1957, pp.107-116.

[7] S.E.G. Öhman, 'Coarticulation in VCV utterances: Spectrographic measurements', J. Acoust. Soc. Amer., 39, 1966, pp.151-168.

[8] M.,Dorman, M. Studdert-Kennedy & L. Raphael, 'Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context sensitive cues', Perception and Psychophysics, 22, 1977, pp.109-122.

[9] K.N. Stevens & S.E. Blumstein, 'Invariant cues for place of articulation in stop consonants', J. Acoust. Soc. Amer., 64, 1978, pp.1358-1368.

[10] L. Lisker & A.S. Abramson, 'Some effects of context on voice onset time in English stops', Language and Speech, 19, 1967, pp.1-28,

[11] A.K. Datta, N.R. Ganguli, S. Ray & B. Mukherjee 'Computer recognition of plosive speech sounds', IEEE Conference on Computers, Session on Pattern Recognition and Learning Methods, 1978, pp.122-134.

[12] P. Alinat, 'Edute du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG', Textes des exposes de 9emes Journees d'Etude sur la Parole, Lanion, France, May-June 1978, pp.297-303.

[13] C.L. Searle, J.Z. Jacobson & S.G. Rayment, 'Stop consonant discrimination based on human audition', J. Acoust. Soc. Amer., 65, 1979, pp.799-809.

[14] H. Fujisaki, H. Tanaka & N. Higuchi, 'Analysis and feature extraction of voiced stop consonants in Japanese', Trans Committee on Speech Research, Acoustics Society of Japan, No. S79-12, May 1979, pp.89-96.

[15] C.J. Weinstein, S. McCandless, L.F. Mondshein & V.W. Zue, 'A system for acoustic-phonetic analysis of continuous speech', IEEE Trans Acoust Speech Sig. Proc., Vol.ASSP-23, 1985, pp.54-72.

[16] W.A. Woods (Principal author), 'Speech understanding system final report', BBN Report No.3438, November 1974-October 1976.

[17] P. Demichaelis, R. De Mori, P. Laface & M. O'Kane, 'Computer recognition of plosive consonants using contestual information', IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-31, 1983, pp.359-377.

[18] M. O'Kane, 'Making the Locus Theory useful for automatic speech recognition', Proceedings of the Tenth International Congress of Phonetic Sciences, edited by A. Cohen and M.P.R. Van den Broecke, Foris Publications, Dordrecht, 1984, pp.331-337.

[19] M. O'Kane, 'Acoustic-phonetic processing for continuous speech recognition', Ph.D. Thesis, ANU, Canberra, 1981.