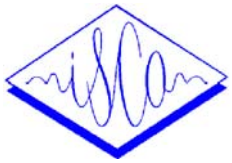
ISCA Archive
http://www.isca-speech.org/archive



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PSEUDO-ARTICULATORY REPRESENTATIONS: PROMISE, PROGRESS AND PROBLEMS.

W.H.Edmondson, D.J.Iskra & P.Kienzle.
Cognitive Science ResearchCentre, School of Computer Science
The University of Birmingham,
Birmingham B15 2TT, UK.
w.h.edmondson@cs.bham.ac.uk

ABSTRACT

Pseudo-Articulatory Representations (PARs) have been proposed and discussed in relation to speech processing with results reported for both synthesis and recognition. PARs are derived from linguistic specifications of articulatory activity which are both abstract and idealized. The abstractions and idealizations permit the linguistic generality to be distinguished from the articulatory reality; this is what we need in speech processing. PARs attempt to retain the linguistic generality whilst also gaining some realism through adoption of continuous articulatory feature values; the latter permits mapping to acoustic values. PARs promise a way of mapping acoustic (and other) data onto linguistic specification of speech (and vice versa), and although enough progress has been made to demonstrate the concept various problems remain for further study.

INTRODUCTION

Pseudo-Articulatory Representations (PARs) have been proposed and discussed in relation to speech processing with results reported for both synthesis [2,5,6] and recognition [7,8]. Our purpose here is to clarify the conceptual issues in order to make PARs more widely accessible to the speech processing community. We also identify and discuss some of the problems which remain to be addressed.

Derivation of PARs

PARs are derived from linguistic specifications of articulatory activity. The insight is that the vocal tract is used for speech in a special way – irrespective of speaker or language. The phonologist views speech as a succession of vocal tract configurational specifications produced with acoustic intent. Conventionally these are thought of as segments, but this too is an abstraction and an idealization. We can take a step towards articulatory reality if we consider these specifications to be targets. We retain the term vocal tract configuration/specification for the linguistically motivated descriptions of what goes on in speech – the succession of targets with acoustic intent; the contrasting term is articulatory configuration/specification which is concerned with the actual physiology.

The vocal tract configurations available for linguistic use are specified in terms of values for each of a set of distinctive features. The specifications are abstracted from the actual activity in the sense that the set of distinctive features is more or less arbitrary. This is saying no more than that there are several different ways of describing/specifying a vocal tract and what can be done with it. Whilst the notion of patterns of distinctly different articulations is quite old [4, cited in 3] the featural basis of specification has been developed considerably, especially in recent times (the familiar table of binary distinctive features was developed in 1952 [9, 1]). A glance at a few text-books will show general agreement on most features, but not unanimity for a full set.

The distinctive features which specify the vocal tract for the linguist in the binary model are also idealised. The model supposes that a specification is one thing or another, but nothing in between. Thus, a sound is either voiced or not (which seems straightforward but in reality may not be) or the tongue is high or not (and low or not, with a mid-position being neither high nor low). But for tongue height, or lip-rounding, or tongue body front/back position, as well as for some other features, the notion of binary values is obviously imposed. This idealization is justified on the grounds that within the overall set of possible configurations the system of binary distinctions accounts for the production and perception of distinctly different speech sounds. This can be checked, within a language, using 'minimal pairs' of words to find meaningful differences between two words differing only in the production of one sound, and there only on the basis of one feature value (the 'minimal' difference). For example /bit/ and /pit/ differ in the value for voicing of the first sound, providing confirmation of the [±voice] distinctive feature. In fact meaningful differences are not essential. Allophonic patterning may provide evidence for the existence of features, as seen for example in the feature [±aspiration] in English where the /p/ is aspirated in /pit/ and not in /spit/. The binary values can be seen to be approximations or idealizations in many cases, for example the lack of aspiration might be set against values of aspiration which are not, in any instrumental sense, 'full aspiration' but rather just enough to provide the contrast. Likewise, to say that the /b/ of /bit/ is voiced and the /p/ of /pit/ is not idealizes the fact that the degree of voicing during the closure of /b/ can vary considerably and indeed even be absent without changing the percept of /b/.

The phonologist's view of articulatory activity, therefore, is abstract and idealized. However, this view does permit the linguistic generality to be distinguished from both the articulatory reality and the acoustic reality. Thus, my tokens of "good morning" will differ from each other, and from the next person's, but the abstraction/idealization permits them all to be categorized as tokens of the type GOOD MORNING. This cannot be overstated – the linguist's insight permits speaker and language independent specification of the vocal tract

10.21437/Eurospeech.1999-370

configuration during speech. This must be so, else we could not understand one another, for the fact is that we do differ in the production of tokens (both from time to time by a single individual, and across a population). However, the linguist's vocal tract specifications look very much like articulatory specifications, and this can be exploited. It is plausible that some abnormal speech habits (e.g. plunging the tongue-tip down behind the lower teeth to involve a different part of the tongue for closing against the alveolar ridge) reveal the hearer/speaker's exploitation, but our concern here is with exploitation as linguists and speech scientists.

PARs are derived from the phonologist's specifications by identifying a sub-set of distinctive features presumed to be especially significant and assigning them values on a continuous scale from 0 -> 1 (or 0->100%, equivalently). This has the effect of mapping many binary features onto somewhat fewer continuously variable features; in both cases the assumption is that the vocal tract configuration targets for linguistic use can be specified in the multidimensional space defined by abstract features. The specification looks articulatory so we refer to the representation of vocal tract targets in these terms as Pseudo-Articulatory Representations.

The utilisation of PARs requires that distinctive feature based accounts of vocal tract configurations (read from textbooks, for example) are matched with evidence of the actual activity (typically this is acoustic). This is considered next.

THE PROMISE OF PARs

Theoretical issues

The first step in exploiting PARs is to map the pseudo-articulatory space (PAR-space) onto some representation of the acoustic signal. This is what listeners and speakers must do, in some strongly equivalent sense, and this observation motivates the claims made below, even where the details of the mapping are not yet known. For example, Iles [6] took formant frequencies, amplitudes and bandwidths as representative of the speech signal (voicelessness notwithstanding) and mapped this space onto the PAR-space (through a set of equations). That this is possible should not cause surprise – the linguist's insight, and thus the PARs, are not concerned with the fine detail of the articulation. For example, I can contact the alveolar ridge with the tip of my tongue for /t/ etc., or with the blade of my tongue a little way back from the tip. The vocal tract 'events' are the same (the linguist's insight) and this is captured pseudo-articulatorily. However, there are several ways in which the linguist's abstractions, as reflected in PARs, ignore potentially relevant aspects of real articulatory behaviour. These are not necessarily problematic, but they do imply that the PAR-based approach needs care.

For example, note that an additional aspect of the linguist's abstraction and generalization concerns differences between languages. The actual articulatory excursions mapped onto the binary features will probably be language specific (the linguist's idealization concerns contrasts not actual values). Because we have acknowledged continuous variations in PAR features we should expect to find language specific differences ("foreign accents") in the use of PAR-space. The implication is simply that language independence in the general model implies some normalization in the use of PAR-space. In fact this is required anyway, to deal with other factors such as speaker characteristics.

Real articulations are variable for non-linguistic reasons, notably differences between speakers. The ventriloquist problem in speech processing (lack of one-to-one mapping between articulation and sound) is actually the solution – we should relinquish the desire to know exactly where each bit of the tongue is because all we need are the gross details (captured pseudo-articulatorily). We are not claiming here that listeners and speakers are unable to process utterance specific or speaker specific details – we can hear someone smiling in a telephone conversation and we can recognize speakers. The promise of PARs is that they represent the linguistic detail only – they convert a vocal tract specification into an articulatory target (and vice versa). Other useful details of the real articulation are not captured or expressed because they are not linguistically controlled. However, we must note that some aspects of speaker characterization may be specified in linguistically relevant terms, and thus in PAR-space. An individual speaker can normalize their [50% round], say, to be a much higher value than the average for a group of speakers.

In general we see the use of PARs as providing both a linguistic handle on the acoustic evidence of articulation, and a way of isolating for other treatment aspects of speech behaviour which are relevant (e.g. speaker identification) but not linguistic.

Practical applications

In addition to the obvious application of PARs in speech synthesis, and speech recognition using acoustic input it is possible to look further afield to possible applications where PARs provide a suitable intermediate representation. One such domain is speech recognition using non-acoustic input.

Speech recognition research is bedevilled by variability in the acoustic input. The elusive goal of free field microphone use in typical office environments – for example, for speech interfaces in Human-Computer Interaction – is widely pursued. One solution to the problem of degraded acoustic input is to augment this input with a non-acoustic source. For example, a Laryngograph could usefully supplement a microphone signal leading to a more robust speech processing front end. Such solutions tend to have the disadvantage that users must be wired up. Other sources of supplementary information would be attractive.

New technology has been developed at Lawrence Livermore National Laboratory, University of California, which promises wire free non-acoustic assessment of vocal tract configurations. The technology is based on a recent development in radar – Micropower Impulse Radar (MIR) [11] Currently heart motion is detectable, and vocal cord motion can be resolved if the radar is held close to the throat. Unconfirmed reports suggest that three radars placed around 1 metre from the head can be combined to provide vocal tract configuration

information (John Holzrichter, personal communication, see also [12]). The power of these radars is very low and the health hazards are negligable.

Information available about the operation of the MIR devices is sparse, but the web pages contain the following observation:

"The microradar can be time multiplexed between two or more range cells so one radar can provide multiple simultaneous outputs, each output corresponding to a different motion sensitive region. When the radiated waveform is sinusoidal, pairs of motion sensitive regions can be offset by 1/4 wavelength so quadrature range-gated Doppler signatures can be obtained for vector processing into magnitude, speed and direction. The multiplexed mode provides motion sensitive slices in depth, so for example, heart motion can be simultaneously recorded at multiple depths into the chest."

The importance of PARs in this application is that they could provide the intermediate representations for both radar derived detail of articulatory behaviour, and acoustically based information, making possible a single representation which provides for recognition on a linguistic basis, as already discussed. The radar based PARs would have to be developed using mappings specific to that particular source of data, but the PARs given as output would be independent of input source (one could not tell from looking at a PAR whence it came).

The value of this example is not its wackiness; rather it shows that PARs promise a practical utility beyond the immediate concern with conventional notions of speech processing.

PROGRESS WITH PARs

Both Iles [2,5,6] and Iskra [7,8] have made use of PARs to explore their feasibility for use in speech synthesis and recognition. Whilst Iles worked with formants, Iskra has worked with cepstral coefficients. The work has demonstrated the value of the general approach, with Iles managing to show recognition feasibility using formants (with specially selected utterances, of course) in addition to synthesis, and Iskra demonstrating some synthesis in addition to recognition. Realistically, however, the results amount to little more than proof of concept, and much remains to be done. In fact, the discussion of problems (below) constitutes a specification for the next phase of the work.

PROBLEMS WITH PARs

The first issue to be addressed is that of the selection of distinctive features as the basis of PARs. The features used ([±high], [±back], [±round], [±tense]), were given continuously variable values and this has worked quite well, but it is now necessary to derive the PARs more thoughtfully. In particular, the feature [±tense] could be considered to be an unobvious choice, because it is not intuitively articulatory in nature (which draws attention once again to the fact that the set of features used by linguists is not in final form). Other features which are more obviously articulatory, and thus better suited to a PAR approach, should be tried. For example the feature [±sonority] is of interest, not least because it also has a range of values (it is inherently non-binary, despite its place in sets of features). The need to work with continuously variable data prompted the move to use continuously variable values for the PARs – but this too needs re-examining. In short, despite the progress made thus far, the basic idea needs reworking on a more principled basis.

PARs are interesting because as approximate accounts of articulatory activity they seem to carry the 'informational burden' in a speaker independent and language independent way, as it were. However, even as an approximate articulatory model it is possible that the pseudo-articulatory space of PARs is still rich enough to account for details which could be speaker or language dependent. Thus, work needs to be done to look carefully at what aspects of PARs provides or underpins the generality we seek (and believe to be present) for both synthesis and recognition. Conversely, it may be possible to use some aspect of the PARs to capture speaker specific information (for example, the "foreign accent" characteristics mentioned earlier).

Another issue in this work is the representation of the speech signal; should we use cepstral or mel-cepstral co-efficients, or formants etc., in isolation, or combination, or what? Additionally, we need to be able to use data on fundamental frequency (for both synthesis and recognition) because although such data generally are not part of the PAR model specifically, they contextualize the PAR data. For example, some information about vocal tract size is available from the fundamental frequency range used by a speaker. Additionally, at least one aspect of fundamental frequency data impinges directly on the PARs and this concerns voicing in the so-called voiced stops. This raises the issue of the need for fine timing details in the pseudo-articulatory model, and these must be supported (this is equivalent to devising a new feature). Note, however, that these details do not turn the model into a realistic articulatory model, they just add the significant details without seeking to model every muscle twitch or glob of saliva.

One problem surfaces quite differently in the two domains – recognition and synthesis. This is the segmentation versus time-sampling problem. In synthesis the targets are determined linguistically. Segments could be used for this, but do not have to be (distinctive features are not predicated on segmentation, although this is often assumed). In fact, it is an attractive property of PARs that non-segmental approaches [10] can be realised as readily as segmental ones. However, having derived a sequence of targets from some sort of input representation, one can set a succession of targets with temporal spacings to suit whatever algorithms one deploys (for prosody, phrase-final lengthening, whatever). A physiologically plausible smoothing algorithm fills in the gaps.

But, in recognition there remains uncertainty even about where the speech begins, and thereafter of course about where it is going. Whatever sampling interval is chosen there will be problems of mapping the data onto putative PAR targets, and dynamic time-

warping, at the least, will probably need to be deployed. However, other sources of structural data are probably available (prosodic data and syllable rhythm data derived from sonority measures) and these may need to be used to control both the sampling and the identification of PAR targets, helping to avoid the spurious recognition of transient articulations as targets. The goal must be a recovered sequence of vocal tract configurations (which could be given as segments). Problems of evaluation can be caused if the to be recognized data are incorrectly labelled and/or are labelled segmentally (for example, the databases currently used in speech recognition work presuppose instantaneous transitions between unambiguously labelled 'segments').

CONCLUSIONS

Some of the above mentioned problems are discussed in Iles [6] and Iskra [8] but it remains the case that they serve to specify the research needed to take the proof of concept closer to a working system. The value of the PARs is that they bridge the linguist's view of vocal activity and the articulatory and acoustic reality. Further, they do so in a way which promises a linguistically sensitive approach to speech synthesis and recognition.

REFERENCES

[1] Cherry, E. C., Jakobson, R. & Halle, M. 1953. Toward the Logical Description of Languages in their Phonemic Aspect. Language, 29:34-46.

[2] Edmondson, W.H., Iles, J.P., & Iskra, D.J. 1996. Psuedo-Articulatory Representations in speech synthesis and recognition. ICSLP'96, 4:2215-2218.

[3] Fromkin, V. A. & Ladefoged, P. 1981. Early Views of Distinctive Features. In Towards a History of Phonetics, eds. R. E. Asher and E. J. A. Henderson. Edinburgh: The University Press.

[4] Holder, W. 1669. The Elements of Speech. London; facsimile reprint, Scolar Press, London, 1967.

[5] Iles, J.P., & Edmondson, W.H. 1994. Quasi-articulatory formant synthesis. ICSLP'94, 3:1663-1666.

[6] Iles, J.P. 1995. Text-to-Speech Conversion using Feature-based Formant Synthesis in a Non-Linear Framework. PhD thesis, Birmingham University.

[7] Iskra, D.J., & Edmondson, W.H. 1998. Feature-based Approach to Speech Recognition. ICSLP'98.

[8] Iskra, D.J. 1999. In preparation. PhD thesis, Birmingham University.

[9] Jakobson, R. 1962. Selected Writings Vol. 1. Mouton & Co.

[10] Kaye, J. 1989. Phonology: A cognitive view. New Jersey: Lawrence Erlbaum Associates.

[11] URL (no date – accessed 04/99): http://www-lasers.llnl.gov/lasers/idp/mir/files/MIR_info.html#organ motion

[12] URL (no date - accessed 04/99): http://www.llnl.gov/das/das_admin/abstracts/abstract3.html