PSEUDO-ARTICULATORY REPRESENTATIONS IN SPEECH SYNTHESIS AND RECOGNITION

William H. Edmondson, Jon P. Iles and Dorota J. Iskra

Cognitive Science Research Centre, School of Computer Science,
The University of Birmingham, Birmingham, B15 2TT, UK.

ABSTRACT

Pseudo-Articulatory Representations are increasingly being used in work on speech synthesis and recognition. The value of such representations lies in their derivation from linguistic abstractions – they are based on articulatory idealizations used by linguists to describe speech. Iles [4] has demonstrated that using these representations it is possible to overcome the many-to-one problem in mapping articulatory configuration to acoustic signal. In this paper we show how the representations facilitate the details of speech processing, for both synthesis and recognition, and we give details of work in progress on recognition. The role of Pseudo-Articulatory Representations in the development of an integrated approach to synthesis and recognition is also discussed.

1. PSEUDO-ARTICULATORY REPRESENTATIONS

The use of Pseudo-Articulatory Representations (PARs) in speech processing has been increasing in recent years. This is true of work in synthesis [6] and recognition [1,2,5]. The motivation for the use of PARs is variable, but in all cases it appears that the main motivation is effectively that such representations constrain some aspect of the processing to fall within physiologically plausible bounds. Whilst this seems to be the general theme it remains the case that sometimes the detailed motivation comes from consideration of problems with processing and sometimes from elsewhere - for examples Iles and Edmondson [4] have been concerned to ensure that speech synthesis is driven in a linguistically plausible manner.

The work reported here is concerned with an integrated system for speech synthesis and recognition, where the integration is made possible by the use of PARs and where the results to date [3,4] demonstrate an important insight in linguistic theory. This work shows, therefore, that the conceptual value of PARs is perhaps only now becoming apparent.

PARs, in the general case, are mappings between properties of the speech signal and parameters with physiological and/or linguistic plausibility. Their value lies in the fact that constraints on values taken by a PAR can be motivated by physiological or linguistic factors. In reality, of course, PARs are mappings between articulatory or linguistic parameters and parameters used to generate speech (e.g. Klatt parameters), or which are derived from speech (e.g. formant values [3]). The constraints provided via this mapping ensure that synthesis is sensibly controlled and that recognition yields plausible values. But there is more to be gleaned from PARs.

Consider the case of recognition (the focus of much of this paper). It has been demonstrated [3,4] that in a simple case, and using PARs mapping formants to modified distinctive features taken from phonology (minor modifications replace some binary features with continuously variable values), it is possible to overcome the *ventriloquist effect*, where acoustic evidence from many different articulatory configurations is recognized as a single phone (see below).

1.1. Linguistic Insight

The significance of this finding (being replicated and extended by Iskra) is that the linguist's distinctive features are ideal abstractions from articulatory reality - *as the linguist would claim*. The abstractions are also ideal in the processing domain because they ensure that the mapping between linguistic representation and signal preserves only those factors which are important. The variation introduced by alternative articulatory detail is rendered insignificant (as it is in fact to the average listener and the linguist).

This observation is really rather important. The implication is that by avoiding the actual detail of articulatory configuration, the processing can avoid dealing with unnecessary articulatory detail which anyway then has to be mapped onto linguistic activity (difficult because of the ventriloquist effect). The use of PARs which are chosen to map the signal to the linguist's distinctive features side-steps an issue in recognition (recognition of irrelevant articulatory detail) and provides a direct path to linguistic representation.

When synthesis is considered alongside this it becomes clear that appropriate choice of PARs provides a route to an integrated system in which linguistic representations in an appropriate phonological format can couple directly to the speech signal - via the PARs which map 'distinctive features' to synthesizer parameters, and (other) PARs which map speech signal properties (formants, LPC co-efficients, etc.) to 'distinctive features'. The processing convenience of PARs is that in fact the processing can be constrained to speech itself, not just a signal processor's idealization of speech.

The detail presented below summarizes already published work on synthesis [4] and shows how the ideas outlined above relate to existing work in recognition, and to our planned work.

2. FEATURE-DRIVEN FORMANT SYNTHESIS

Feature-driven Formant Synthesis (FDFS) was designed to test the feasibility of producing synthetic speech using a formant

synthesizer driven by articulatory style features. The second goal was to be able to synthesize speech from phonetic input.

In the initial stage, a set of phones and a set of articulatory features were selected [3,4]. The latter were represented by distinctive features whose binary values were in many cases replaced by continuous ones as the continuous representation seemed more adequate and appropriate to describe natural speech. That is why these features will be referred to from now on as *quasi-articulatory features* and can be considered as a linguist's abstract view of articulation. The final requirement was a set of synthesizer parameters needed to produce a synthetic version of each phone. The Klatt synthesizer design in the parallel only mode was selected and the parameters were extracted from natural speech. Both the quasi-articulatory features and the relevant synthesizer parameters were fed into multiple regression analysis to provide a mapping between the two sets. Now that all the data were available to produce discrete speech segments, a technique was required to enable the synthesis of continuous speech. A cosine interpolation was used to transform the segmental articulatory descriptions into continuous trajectories. Finally, the Modified Rhyme Test was used to evaluate the quality of the synthesized speech.

One of the advantages gained by using articulatory controls to drive the synthesis process is the ability to vary the precision with which the synthetic speech is articulated. This is done by manipulating the degree to which the articulatory trajectories achieve the idealized targets and can be used to model rapid speech [4].

2.1 Inverse Mapping

Another advantage offered by FDFS is simplification of the inverse articulatory-acoustic mapping problem. The problem with articulatory-acoustic mapping is that it is a many-to-one mapping, better known as the ventriloquist effect, where it is possible for a number of different articulator configurations to produce a particular set of acoustic cues, and these acoustic cues are recognized as a particular phone.

Initially, the inverse mapping was attempted from perfect data, namely the Klatt parameters. Using a brute search tool an output was generated describing the trajectories of continuous features. The assumption was made that all the input speech was voiced, so that only a restricted feature set would be relevant. The search was constrained by applying weights to emphasize the lower formants and restricting the distance measures between the formant frequencies. Additionally, the scope of variation for the movement of each articulator was constrained as they are relatively slow-moving in reality. Resynthesis produced convincing speech.

Having succeeded in deriving features from synthesized speech, the inverse mapping was performed using formant data extracted from *natural* speech. In order to compare the old and the new formant trajectories, the latter were used as input to the FDFS model to produce a synthetic copy of the original utterance. The results of the inverse mapping were felt to be satisfactory and informal listening to the copy-synthesis revealed a close match to the original synthetic speech.

The inverse mapping technique described above can be used to extract feature trajectories from natural speech to be used for recognition. In order to achieve that, the inverse mapping has to be extended to voiceless speech as well. This is the direction the research on this project is going to take in the near future.

3. SPEECH RECOGNITION BASED ON FEATURE EXTRACTION

Current research concerning use of distinctive features in speech recognition is not very extensive. Johnson [5] models speech recognition as the estimation of distinctive feature values at articulatory landmarks. Each distinctive feature is modelled as a table containing phonetic contexts, acoustic correlates for this feature in each context and for each context a statistical model for evaluating the feature given the measurements. He claims distinctive features are superior to phonemes since there are fewer of them and they can represent coarticulation more adequately. He also claims that the linguistic content of speech is located at a series of articulatory landmarks and the speech signal in between carries information about feature values at these landmarks.

In the recognition process the values of a feature are estimated as the recognizer reads through the list of possible contexts until a match with the current landmark is found. Then the corresponding acoustic correlates are measured and their values are used to classify the feature.

The problem with this approach lies in the fact that concentrating on prominent acoustic landmarks may lead to overlooking a great deal of acoustic detail which may be helpful in recognizing certain sounds.

Deng and Erler [1] use phonetic features in a different way: they employ them as the basic speech recognition units which are used to train Markov models. A set of multivalued phonetic features with strong acoustic and articulatory correlates are chosen. Each state represents a combination of feature values and each combination represents, in turn, a stable acoustic property resulting from a particular articulatory configuration. Coarticulatory effects are incorporated by allowing asynchronous time alignment of different features over adjacent phones, which is different from the conventional segmental model.

Deng and Erler build a phonetic-features model which is based on a phonetic-feature space and a vocabulary description prescribing a related sequence of target feature vectors for each word. This model is built for a vocabulary consisting of words which are stop consonants followed by a vowel. A composite HMM is constructed by assigning to the free-valued features all possible feature values (restricted only by the monotonicity constraint) and by creating a set of transitional states so that any sequence of transitions through the model corresponds to monotonic changes of feature values in any feature dimension.

In the recognition task several sets of HMMs are trained using different modelling units, such as microsegments, allophones, phones, words and features. The recognizer based on feature models achieves significantly better results than the others. However, the current model has only been trained to recognize

simple consonant-vowel clusters and its complexity is expected to increase geometrically when the vocabulary is extended.

Espy-Wilson [2], too, creates a new approach to speech recognition based on distinctive features. As an alternative to Mel-cepstral coefficients she introduces a new signal representation where parameters are designed to capture acoustic properties of manner-of-articulation features. This approach enables reduction of the cross-speaker and contextual variability and creates an acoustic-oriented rather than segment-oriented approach to speech recognition. In this approach specific acoustic events are identified around which acoustic properties of features are extracted. No assumption is made that sounds are non-overlapping, which makes it possible to deal with the problem of coarticulation.

Espy-Wilson builds a system for recognizing semivowels. She selects a database of words containing semivowels in different environments. Then features are chosen to separate semivowels as a class as well as to distinguish between them. Acoustic correlates are determined based on relative measures of successive parts of the spectrum as opposed to absolute values, which minimizes their sensitivity to the speaker, the speaking rate or the context. Specific events in the appropriate acoustic parameter, such as e.g. a minimum, mark a change in the value of a feature and serve as landmarks for extracting the acoustic correlates for features.

Two sets of HMMs are trained: one based on phones and the other on features. The feature-based recognizer performs better and detects a higher percentage of semivowels correctly.

Despite the interesting way of dealing with coarticulation, this approach leaves too much room for subjective judgements concerning the feature extraction procedure. Moreover, many problems are bound to arise when other classes of sounds, such as obstruents, are tackled.

3.1. Future Work on Recognition

After the first successful attempts at inverse articulatory-acoustic mapping by Iles [3,4], the approach has to be extended to incorporate other classes of sounds. There are two problems connected with this: one concerns obtaining reliable spectral data for all classes of sounds and the other is establishing correlates between the features and the segmental representation.

As far as the first problem is concerned, it is well-known that formant trajectories can only be used to distinguish among different vowels and perhaps semivowels. For other classes of sounds, such as obstruents, it is necessary to employ some other analysis techniques which incorporate all the acoustic detail and base their parameter extraction on the whole spectral envelope. Two ways of describing the spectral envelope are taken into account here: linear predictive coding and Mel-cepstral coefficients.

These techniques should make it possible to extract parameters from the speech signal which will be powerful enough to describe all sounds, not only vowels, and discriminate among

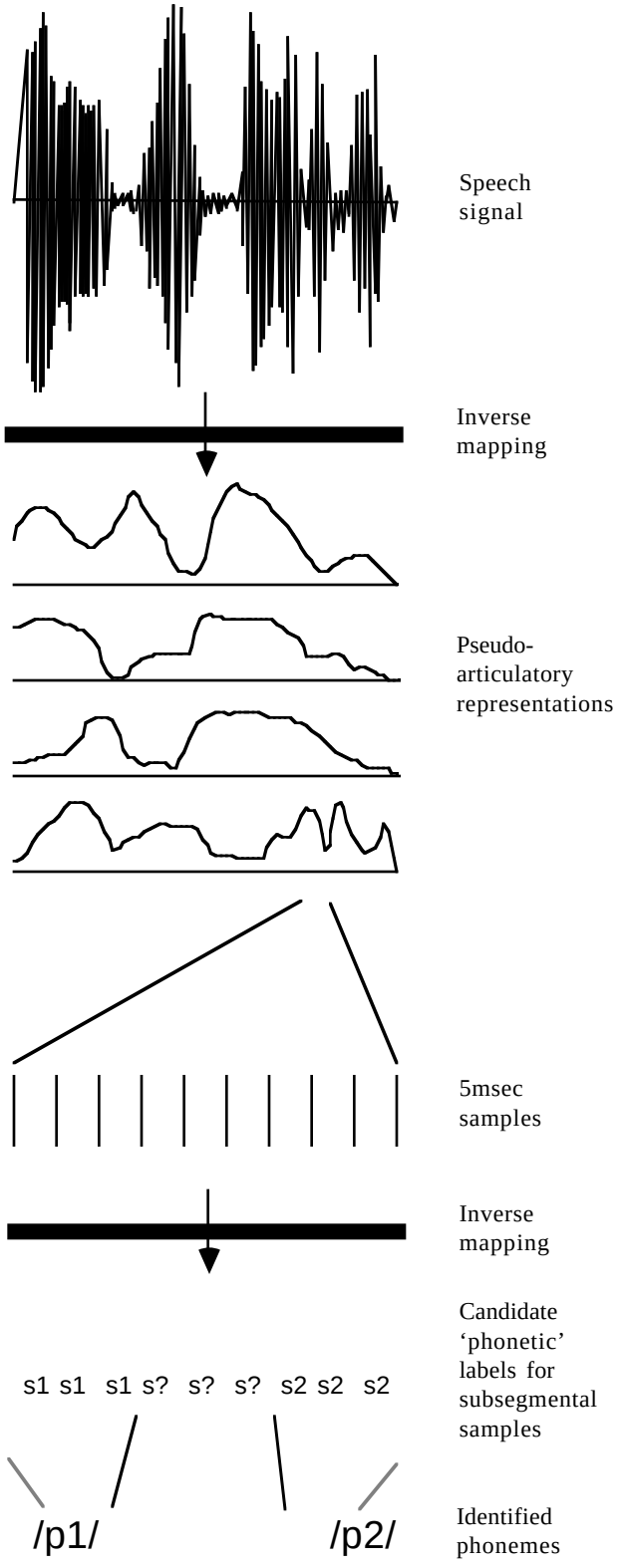

Figure 1 Schematic diagram of recognition process using Pseudo-Articulatory Representations.

them. Experiments will have to be carried out to test which technique is the best. These parameters will serve to obtain PARs using multiple regression analysis again.

Having made the transition to the PAR space, as shown in **Figure 1**, this space will have to be transformed again to obtain a textual representation of the output. Just like in the case of the first transformation, multiple regression analysis will be used to map the features onto segments. To each frame in the feature space a segment label will be assigned. This will yield a large number of labelled sub-segments, with the phonetic segments of interest lasting longer than a frame and comprising several successive occurrences of labelled sub-segments. In order to prevent non-permissible phone strings a simple grammar can be built in to constrain the choice of possible segments.

4. CONCLUSIONS

In addition to repeating the earlier demonstration that the ventriloquist effect can be overcome [3,4], and with a full range of phones, we believe this work offers three important contributions.

Firstly, the frame by frame analysis and conversion to segmentally labelled sub-segments provides a sensible way of handling noisy input. Working at a simple segment scale raises problems caused by errors in identification of segments: segment sequences become errorful in ways which are difficult to detect and correct (although segment transition probabilities can help). Sub-segmental processing provides repeated measures within any one phone and it thus provides for both error correction in phone identification (e.g. the transition probabilities between frames are not the same as those between phones) and it provides the route to explicit identification of segment transition boundaries.

Secondly, the fundamental insight in the linguist's description of articulation in idealized terms is shown to be practically useful – this is what makes it possible to overcome the ventriloquist effect.

Thirdly, the use of PARs clearly opens the way to developing comprehensively integrated systems for synthesis and recognition (see also Iles [3]) which is highly desirable from a modelling point of view. The issue here is that full integration enables system builders, as well as cognitive scientists who wish to develop models, to utilise integrated models and databases for the linguistic and articulatory sub-components. The gain is not just one of efficiency: full integration is appealing cognitively, and ultimately it makes progress toward more complex systems (for example text-to-speech combined with speech-to-text) very straight-forward.

5. REFERENCES

1. L. Deng and K. Erler. Structured design of a Hidden Markov Model speech recognizer using multivalued phonetic features. *J. Acoust. Soc. Am.*, 92, 1992.

2. C. Y. Espy-Wilson. A feature-based semivowel recognition system. *J. Acoust. Soc. Am.*, 96, 1994.

3. J. P. Iles. *Text to speech Conversion Using Feature-Based Formant Synthesis in a Non-linear Framework.* PhD thesis, University of Birmingham, School of Computer Science, 1995.

4. J. P. Iles and W. H. Edmondson. Quasi-articulatory formant synthesis. *ICSLP'94*, 3:1663-1666, 1994.

5. M. Johnson. Automatic context-sensitive measurement of the acoustic correlates of distinctive features at landmarks. *ICSLP'94*, 3:1639-1642, 1994.

6. R. Wilhelms, P. Meyer and H. W. Strube. A quasi-articulatory speech synthesizer for German language running in real time. *J. Acoust. Soc. Am.*, 86(2):525-539, 1989.