Developing a Nonsymbolic Phonetic Notation for Speech Synthesis

Andrew Cohen*
University of Reading

The goal of the research presented here is to apply unsupervised neural network learning methods to some of the lower-level problems in speech synthesis currently performed by rule-based systems. The latter tend to be strongly influenced by notations developed by linguists (see figure 1 in Klatt (1987)), which were primarily devised to deal with written rather than spoken language. In general terms, what is needed in phonetics is a notation that captures information about ratios rather than absolute values, as is typically seen in biological systems. The notations derived here are based on an ordered pattern space that can be dealt with more easily by neural networks, and by systems involving a neural and symbolic component. Hence, the approach described here might also be useful in the design of a hybrid neural/symbolic system to operate in the speech synthesis domain.

1. Background and phonetic motivation

Phonological and phonetic notations have been developed by linguists primarily as descriptive tools, using rewrite-rules operating on highly abstracted basic units derived from articulatory phonetics. Even some connectionist work has followed this tradition (Touretzky et al. 1990). The primary aim of these notations is explanation and understanding, and there are difficulties in incorporating them into systems with a practical aim such as deriving speech from text, which tend to be data-driven. One recent study claimed that introduction of linguistic knowledge degrades performance in grapheme-phoneme conversion (van den Bosch and Daelemans 1993). However, typical purely data-driven systems are opaque from a phonetic or phonological point of view. In order to handle many of the very hard problems remaining in speech synthesis, there is a need to develop a basic underlying notation (or method of deriving a notation) that can be parameterized for different speakers. This notation could be based on articulatory phonetics (where a higher-level task, such as grapheme-phoneme conversion, is being performed) or on a spectral/perceptual measure of similarity, for more low-level tasks such as duration adjustment. This notation would ideally be represented in a low-dimensional, topological space so as to be both perspicuous and flexible enough to use in further nonsymbolic modules.

Existing synthesis-by-rule (SBR) systems (Allen et al. 1987) have been concerned with text-to-speech conversion, and have made use of a segmental approach derived from traditional phonology. Among the simplifying assumptions remaining from this approach are that transitions into and out of a consonant are identical, and that the same transition may be used in each CV combination, regardless of the larger phonetic environment. These assumptions need to be modified in a principled manner rather than by tables of exceptions.

* Department of Cybernetics, University of Reading. E-mail: cybadc@cyber.reading.ac.uk

It has been argued by phoneticians that articulatory models cannot account for all the variability found in natural speech (Bladon and Al-Bamerni 1976; Kelly and Local 1986). Therefore, there is a need to find ways of incorporating other sources of variability into synthetic speech, including, for example, the feedback a talker receives from the perception of their own voice. Evidence that such feedback affects speech is the degradation seen in the speech of persons with acquired deafness. One possible way to introduce this kind of variability is through the development of representations that encode (in a reduced dimensionality) a range of examples of the phenomenon to be accounted for. Formant data can be used to introduce a perceptual measure of similarity (see section 3 below).

This report describes the theoretical motivations of an experimental system that has been implemented as a set of shell scripts and 'C' programs; not all of the technical details of this system have been finalized, and it has not been formally tested. While formants have been made use of as training data (as well as acoustic tube data), as yet no use has been made of a formant synthesizer for creating the output speech, due to the need for handcrafting of values. At present, waveform segment concatenation is being used to explore a parametric duration model based on the kind of proximity-based notations described here.

2. Application of the SOM to phoneme data

In outline, the Self-Organizing Map (SOM, Kohonen 1988) approximates to the probability density function of the input pattern space, by representing the N-dimensional pattern vectors on a 2D array of reference vectors in such a way that the resulting clusterings conform to an elastic surface, where neighboring units share similar reference vectors. This algorithm and Learning Vector Quantization (LVQ) are described in Kohonen (1990), which has practical advice for implementation, and in more theoretical detail in Kohonen (1989).

It has been widely noted that 2D representations of speech are useful where there is a need to transmit information to humans at a phonetic level—for example, in tactile listening systems (Ellis and Robinson 1993). If a speech synthesis system has a phonetic interface or level of operation, it is then possible to introduce learning techniques for subsequent modules (e.g., those which calculate durations or an intonation contour) and to have an idea of what is happening, in phonetic terms, when things go wrong, and therefore how the training program or learning method may be adjusted. There is a long tradition of two-dimensional representations of formant data in attempts to classify vowels, going back at least to the study of Peterson and Barney (1952). Another type of advantage lies in the flexibility given by the very large dimensionality reductions achievable by Kohonen's technique. These reductions are possible even where the input pattern space may be only sparsely populated, yielding a flexible encoding with not too many degrees of freedom. It is possible for Kohonen's technique to work in 3D (3D maps have been produced by the author, but are more difficult to work with and are still undergoing evaluation). In 4D or above, interpretation becomes much more difficult. Refinements such as the Growing Cells technique (Fritzke 1993) might be preferable to a move to higher dimensionality, so as to retain transparency of the notation and a possible link to symbol-based stages of operation.

Figure 1 shows a map resulting from applying the SOM algorithm to phoneme feature data. The following nine binary articulatory features were used: continuant, voiced, nasal, strident, grave, compact, vowel height(1), vowel height(2), and round. The features h1 and h2 are used for height simply because there are three possibilities:

```
f    .    s    .    sh   .    ch   .

.    .    .    .    .    .    .    j

v    .    z    .    dh   .    .    .

.    .    .    .    .    th   .    nch

r    .    l    .    .    .    .    .

.    y    .    .    m    .    nk   .

.    .    .    n    .    b    .    g

w    .    .    .    .    .    .    .

.    u    .    i    .    d    .    .

.    .    .    .    .    .    .    .

o    .    e    .    t    .    p    .

.    @    .    a    .    ^    .    k
```

Figure 1
Clustering of phoneme data (8 × 12).

open, mid and closed, which cannot be encoded by a binary bit.[1] In this case, the point is not to do feature extraction (since the features are already known), but to provide a statistical clustering in 2D that can indicate whether the features chosen provide a good basis for analysis. Figure 1 suggests that phoneticians have 'got it right' in that the features do result in a clustering of similar sounds such as stops, fricatives and nasals, as well as the more obvious separation between vowels and consonants. It is worth pointing out that neither the SOM nor the LVQ algorithm handles raw data (such as waveform values or image intensity values), but each operates on data such as spectral components or LPC coefficients that are themselves the output of a significant processing stage, and can justifiably be called features.

The phoneme map is produced by a single Kohonen layer that self-organizes using the standard algorithm (Kohonen, 1990), taking as input nine articulatory features commonly used by phoneticians to describe the possible speech sounds. The features were designed so that any phoneme (or syllable) may be uniquely specified as a cluster of features, without reference to specific units (segments such as phones, syllables, etc.)—any feature may run across unit boundaries. Figure 1 shows a 12 × 8 map created (as are all the following maps) with hexagonal connections in the lattice indicating which units are neighbors. A monotonically shrinking 'bubble' neighborhood was used in all the maps shown here. Kohonen refers to this type of kernel as a bubble because it relates to certain kinds of activity bubbles in laterally connected networks (see Kohonen 1989).

[1] Thanks to John Local for providing the basis for the data.

The analysis of these maps is at a phonemic level of description; this is a very abstract level compared to the phonetic descriptions typically used, which take into account much more of the context. However, the trend in recent phonology has been towards ever greater abstraction and more complex hierarchies of units (e.g., Goldsmith 1990; Durand 1990). The abstractness of phonemes makes them more difficult candidates for both recognition and synthesis, although most existing systems in the two fields perversely make use of a phoneme stage. If a phoneme stage is held to be essential (on grounds of parsimony, perhaps), then maps of the type shown in figure 1 may be one means to incorporate phonemes into a static or recurrent neural network-based system in a more flexible fashion. However, the essential point is that trajectory across the map provides a bridge between the symbolic description of the data and the data itself. Robustness of mappings between domains (e.g., from text to phonemes) should be increased, since similar sounds (words) will have similar trajectories across the map.

Clearly, it would be interesting to repeat the process with formant data to see if a similar 2D map can be formed. Formants are known to be important in perception, but do not in general correspond to a particular vocal tract configuration. Many papers have stressed the importance of formant peaks in speech perception, especially in the case of vowels (Klatt 1982a). Changes in formant frequency are more important in phonetic judgments than changes in formant amplitudes or bandwidths, or zeros in the spectrum (Klatt 1982b).

The diphone approach embodies a signal-modeling, as opposed to a system-modeling, viewpoint. Therefore, it is possible to incorporate sources of variability other than the purely articulatory into the learning procedure that derives the basic notation. Features such as accent-specific detail, stress and emotional quality are difficult to describe in purely articulatory terms, and, with current knowledge, cannot be given a re-usable, speaker-independent representation at all.

3. Application of the SOM to diphone formant data

The formant data obtained (F1, F2, F3, F4 and AV; formant bandwidths were not used) in each frame were passed as a single vector to the SOM. (For a description of the SOM algorithm see Kohonen 1990, 1989). Some diphones would run over more frames than others, so shorter vectors were padded with zeros to make them up to the length of the longest. The average length of a diphone (unpadded) was about 30 frames; the bulk of this would naturally consist of steady-state rather than transitional data. However, the training sets in the maps shown were chosen so that the significant variance would lie in the transitional parts of the training vector, rather than in the less interesting steady-state portions.

A similar type of experiment to that in section 1 can be carried out on diphone formant data. For comparison, figure 2d shows the same data set as in figure 2c clustered by means of Sammon's mapping (Sammon 1969). The latter is a supervised, error minimization procedure that maps the input vectors onto a lower dimensional space in such a way that the Euclidean distance between the endpoints is approximately preserved. It does not have the topographic property of the Kohonen map, but is useful in indicating the underlying form and orientation of the data in 2D.

A number of maps showing clustering of similar sounds (formant vectors) have been obtained (Figures 2a–c show examples). Larger maps containing up to 1250 diphones have also been created. In contrast to the one-point representations used in Huang (1990) for vowels, the entire diphone is presented to the network for classification. The frame length (not more than 10ms) should mean that all perceptually

.	.	.	.	aar	.	aag	.	aam	.
.	.	aaw	.	.	.	.	.	aap	.
.	.	.	.	aay	.	.	.	.	aak
.	.	aazh	.	.	.	.	aat	.	.
.	.	.	.	.	.	.	.	.	aas
.	aash	.	aaj	.	.	aath	.	.	.
.	.	.	.	aal	.	.	.		.
aaz	aad	.	aah	.	.	.	.	.	.
.	.	.	aaf	.	aang	.	aan	.	aadh
aab	aach	aav	.	.	.	.	.	.	.

Figure 2a
Clustering of diphone data for aa-C.

relevant information is captured in the formant trajectories. Maps based on acoustic tube data computed from the LPC coefficients have also been created, with much the same kind of results as seen in the formant maps. That the results should be similar is to be expected as this data is essentially spectral, and bears little resemblance to real vocal tract data. Experiments are currently being carried out to determine whether these maps or those based on formants will work better as part of a prototype speech synthesis system.

To factor out the influence of the initial configuration of the network (the reference vectors are initialized to small random values), twenty trials were run on each data set, and the map with the lowest quantization error (QE) was selected as the best. The QE is simply the mean error over the N pattern vectors in the set,

$$QE = \frac{\sum_{t=1}^{N} \|x(t) - m_c(t)\|}{N}$$

where $x(t)$ is the input vector and m_c the best matching reference vector for $x(t)$.

In order to compare QEs, the topology (form of lateral connections) and adaptation functions must be the same, since the amount of lateral interaction determines the self-organizing power of the network. In the simplest case of competitive learning the neighborhood contains only one unit, so a minimal QE may be achieved, but in this case there is no self-organizing effect.

Schematically, then, resynthesis would take place on the basis of a trajectory across a diphone map. The trajectory could be stored simply as a vector of co-ordinates that are 'lit up' on the map. These vectors would occupy little storage space, and might be passed as input to a further SOM layer to try to cluster similar sounding words. The time-varying, sequential properties of speech, which are difficult for neural nets to handle, can thus be modeled as a spatial pattern in an accessible and straightforward manner. Vectors of addresses would be completely different (e.g., the endpoints

```
daa  .    aab  .    aap  .    aach aaf

.    .    aag  .    .    aah  .    .

gaa  .    .    aaj  .    .    aas  aash

.    baa  .    .    .    aaz  .    .

.    .    .    aad  aadh .    .    aazh

.    faa  .    .    .    aam  .    .

saa  .    .    paa  .    .    .    aal

.    haa  .    .    .    aay  .    .

.    .    .    jaa  .    .    aan  .

.    .    .    .    .    naa  .    aar

waa  .    zaa  ngaa laa  .    .    .

maa  .    .    .    .    .    raa  .

.    .    dhaa .    .    chaa .    vaa

taa  .    .    kaa  .    .    .    thaa
```

Figure 2b
Clustering of data for aa-C and C-aa.

would be far apart), although the closeness of similar individual segments is maintained in 'diphone space' on the map. Current practical work is focusing on developing a method for determining segment durations based on SOM distances between successive diphones in a string.

The creation of a 'diphone space' in 2D may assist both in choosing the correct (best matching) segments and in interpolation at segment boundaries. In a text-to-speech application, when mapping from text onto diphones, it is clearly an easier task for an MLP to map neighboring regions of input space onto neighboring regions of output space. In joining segments the amount of distance travelled across the map is evidence of how different the two segments are in perceptual terms, and, therefore, of the amount and type of smoothing of formant values needed. The maps also suggest an approach to the problem of calculating segment durations, which is currently under investigation. For example, consider the sequence 'aalaa'. This can be made up of the diphones 'aal' and 'laa', which are close in diphone space. This suggests a shorter duration for 'aalaa' than for 'aalpaa', where there is a transition into a consonant, which is acoustically very different (and further away on the SOM). These considerations are of course subject to modification by prosodic factors, such as the need to stress a particular word.

In the prototype system under development, the baseline system uses a diphone

```
airp.    .    .    ahp  .    aap  .    ep   .       Key:
                                                    Sym Example
.   oap  .    iep  .    oip  .    .    .    op      "air" /air/
                                                    "ah"  /bard/
.   .    .    .    .    .    .    oop  .    .       "e"   /bed/
                                                    "oa"  /oak/
aip oorp .    erp  .    awp  .    .    .    ip      "ie"  /pie/
                                                    "oi"  /oil/
.   .    .    .    .    .    .    uup  .    .       "oo"  /good/
                                                    "ai"  /pain/
arp .    .    pee  .    .    .    .    .    eep     "oor" /poor/
                                                    "er"  /bird/
.   .    .    .    .    pu   .    po   .    .       "aw"  /board/
                                                    "i"   /bid/
poi .    .    paa  .    .    .    .    .    pe      "uu"  /brood/
                                                    "ee"  /bead/
.   .    .    .    .    .    .    .    .    .       "u"   /bud/
                                                    "o"   /body/
poa .    .    .    .    .    pair .    .    pi      "uh"  /above/
                                                    "ou"  /out/
.   .    peer .    .    per  .    .    poo  .       "eer" /ear/
                                                    "ar"  /art/
.   .    .    .    .    .    .    .    .    puh

poor.    .    .    pou  .    .    .    .    .

.   .    paw  .    .    pai  .    puu  .    p-

pie .    .    par  .    .    .    .    .    -p
```

Figure 2c
Clustering of data for transitions into and out of 'p'.

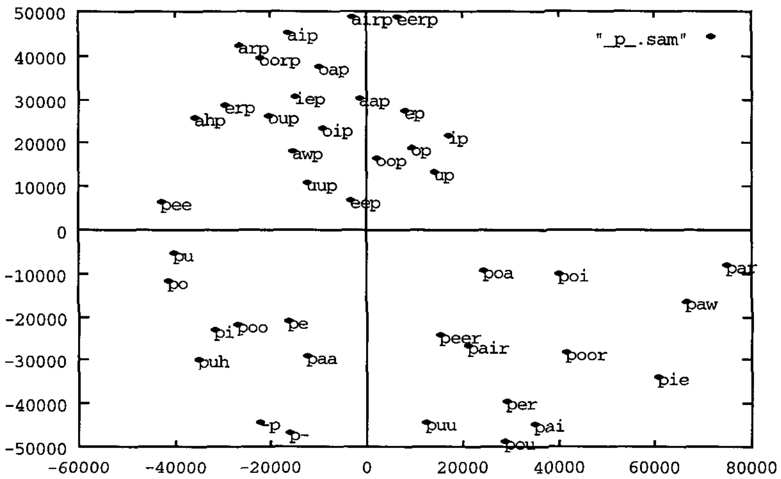

Figure 2d
Clustering of data with Sammon's mapping.

concatenation procedure, on which various enhancements based on the SOM are being tried, which will be more fully described in future reports. Using the examples given on a record supplied with Klatt's (1987) review article, informal comparison shows a high degree of variability in quality of the sentences generated: the best are comparable with the diphone concatenation methods (which have better transitions than DECtalk, even if the prosody is in some cases not as well developed), while the worst are highly unnatural, but usually intelligible.

4. Conclusion and further work

The outline of a conventional SBR system has a series of symbolic stages, assuming a modularity of data at each level, before the final low-level stage ('synthesis routines') calculates the synthesizer parameters. The essential feature is the 'abstract linguistic description', which must be derived before any attempt is made to calculate parameter values. In the proposed system, this middle stage is replaced by the SOM stage, which introduces a learned notation based on acoustic data. Generation of an intonation contour, though this has been implemented with neural nets, is probably best handled with rules as it is almost purely a prosodic (i.e., sentence level) matter.

The SOM coding replaces the linguistic description, and leads to direct access of waveform values for a given diphone, which then become default values for the next stage to operate on. In conclusion, arguments have been presented for the use of nonsymbolic codings as the central stage of a text to-speech system. These codings are both closer to the acoustic domain and capable of greater flexibility than the standard phonetic notations. Additional sources of variability, such as stress and emotional quality, could also be accounted for with this kind of trajectory in a low-dimensional space, rather than attempting to derive a speaker-independent symbolic notation. These maps are also capable of being operated on by a neural network in further processing stages, opening the way to a different type of phonetics based on a multitude of soft constraints rather than the rigid phoneme and rewrite rule.

Further work is needed to investigate the usefulness of the SOMs in speech synthesis, and how they may be integrated in a hybrid system that uses rule-based prosody. Other data sets need to be explored to introduce other kinds of variability. It would also be important to determine whether the distance measure provided by the diphone maps correlates better with subjective perception of the mismatch between successive diphones than more standard measures of spectral distance, such as various distance measures between frames of cepstral coefficients.

Acknowledgments

Thanks to Stephen Isard of Edinburgh CSTR and Linda Shockey of the Linguistics Department, Reading University for help with diphones and related matters. The author is grateful to John Local for enthusiasm and help with phonetics. Any remaining mistakes are the author's responsibility. Thanks to the Laboratory of Computer and Information Science, Helsinki University of Technology for making available their 'SOM_PAK' software, used here with minor modifications.

References

Allen, J., Hunnicutt, M. S., and Klatt, D. H. (1987). *From text to speech: The MITalk system*. Cambridge University Press.

Bladon, A., and Al-Bamerni, A. (1976). "Coarticulation resistance in English." *Journal of Phonetics*, 4, 137–150.

van den Bosch, A., and Daelemans, W. (1993). "Data-Oriented Methods for Grapheme-to-Phoneme Conversion." *Proceedings, 6th Conference of the European Chapter of the ACL, Utrecht, April 1993*.

Durand, J. (1990). *Generative and Non-Linear Phonology*. Longman, London.

Ellis, E. M., and Robinson, A. J. (1993). "A Phonetic Tactile Speech Listening System." Cambridge University Engineering Department Technical Report, CUED/F-INFENG/TR122, May.

Fritzke, B. (1993). "Growing Cell Structures—A Self-organizing Network for Unsupervised and supervised Learning." International Computer Science Institute Technical Report TR-93-026, May.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Blackwell, Oxford.

Huang, C. B. (1990). "Modelling Human Vowel Identification Using Aspects of Formant Trajectory and Context." In *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka. *Proceedings, ATR workshop*, Kyoto, Japan, November 1990, IOS press, Oxford, UK.

Kelly, J., and Local, J. (1986). "Long-domain resonance patterns in English." In *Proceedings IEEE Speech Input/Output Conference*, Pub. No. 258, 77–82.

Klatt, D. H. (1982a). "Speech processing strategies based on auditory models." In *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom, Elsevier, Amsterdam.

Klatt, D. H. (1982b). "Prediction of perceived phonetic distance from critical band spectra: a first step." In *Proceedings IEEE ICASSP-82*, 1278–1281.

Klatt, D. H. (1987). "Review of text-to-speech conversion for English." *JASA* 82(3), 737–793.

Kohonen, T. (1988). "The 'neural' phonetic typewriter." *IEEE Computer* 21, 11–22.

Kohonen, T. (1989). *Self-Organization and Associative Memory*. Springer Verlag, 3rd ed.

Kohonen, T. (1990). "The Self-Organizing Map." *IEEE Proceedings* 78(9), 1464–1480.

Peterson, G., and Barney, H. (1952). "Control methods used in a study of the vowels." *JASA* 24, 175–184.

Sammon, J. W. (1969). "A nonlinear mapping for data structure analysis." *IEEE Trans. Computers*, C-18, 401–409.

Touretzky, D. S., Wheeler, D. W., and Elvgren III, G. (1990). "Rules and Maps III: Further Progress in Connectionist Phonology," School Of Computer Science, Carnegie Mellon, Technical Report CMU-CS-90-138