Self-Organizing Networks relate Phonetic and Articulatory Speech Data

Rik Crabbe Jacques J Vidal Georges Papcun

ABSTRACT

This paper suggests that articulation would assist in phonetic identification and ultimately in automated speech recognition.

INTRODUCTION

The articulatory behavior, i.e., the motion of the organs of speech and vocal tract shape, contains concurrent and separate information about the spoken message. It follows that knowledge, even incomplete, of the articulation trajectories during speech can improve phonetic identification. This fact is directly observable in humans since comprehension can be notably improved with lip reading and facial expression. It has also deep implications regarding the relationship between production and understanding of speech. Recent work in theoretical linguistics suggests that humans do *infer* articulation when they perceive speech (Nittrouer & Munhall 1988) This claim is consistent with the neurological view of the unity of perceptual and motor mechanisms, i.e., that the perceptual recognition of a language coded message and its motor production share common mechanisms.

These claims have implications for the computer understanding of speech . They suggest that articulation measurements, in conjunction with a model of speech production, would assist in disambiguating the acoustic signal and ultimately improve speech recognition. A similar idea is is found a study in the recognition of hand-writing recognition based on the production of the hand motion by one of the authors (Hoffman, Sckrzypek and Vidal, 1992).

Computer experiments have related articulation, i.e., from the speech sound wave (Atal, 1970, Wakita, 1973&1979; Protter, 1987). Recently, one of the authors demonstrated a backpropagation neural network *predicting* articulation from the sound of spoken voyels (Papcun et al., 1992).

This paper builds upon this experiment and on a methodology developed by Kohonen who developed phonotopic *features maps* over which the phoneme sequence of continuous speech would trace two dimensional trajectories (Kohonen 1990). On the maps phonemes emerged in distinct spatial locations and similar phonemes tended to cluster. Furthermore however, the maps showed a striking correspondance with non linear projections of data from a mechanical models of the vocal tract and resonance cavities. In this paper we examine self organizing feature maps that relate articulatory data to speech.

PROBLEM STATEMENT

This paper reports preliminary results in an ongoing research toward validating the assumptions just listed. The intention is to explore the structure of speech articulations with self organizing maps and to evaluate the potential of such knowledge in supporting speech understanding both with or without the benefit of real-time articulatory motion measurements.

DATA

We have used data collected at the Waisman Center of the University of Wisconsin. The data consists of samples of voiced speech and corresponding articulatory trajectories. The gathering procedure is reviewed briefly here. additional details can be found in Nadler et al., 1987 and Papcun et al., 1992.

The acoustic signal was recorded at 10,000 Hz and then later resampled to Codec (8012.821) for use with NeXT hardware. The acoustics were then divided into 50 % overlapping windows 128 samples wide. The articulatory data was collected with a 1 mm diameter x-ray that sampled the positions of several 3 mm diameter gold pellets affixed to the subject's head, lips and tongue. The microbeam rotated between the pellets, sampling the slower articulators at 90 Hz and the faster articulators at 180 Hz. A pellet's position, velocity and acceleration at one sample was used to aim the beam for the next sampling. The articulatory trajectories were then resampled at 125.2 Hz using a cubic spline (Press et al., 1988) to place each sample at the center of its acoustic window.

The study used 4 subjects, 2 female and 2 male. Each subject was asked to read several word lists as naturally as possible. The lists varied slightly from subject to subject, but all 4 read approximately 135 words. These words lists covered a broad range of english phonemes and phoneme combinations. Phonemes identified by hand for all data from one male subject were later used to label the networks. Because of the difficulty of finding divisions between vowels and liquids, certain phonetic digraphs were used such as /a'/ and /o'/.

NETWORK INPUT VECTORS

Overlapping windows , each containing 128 raw samples were FTT transformed to provide a input representation based on frequency domain variables. To obtain input vectors for the neural networks, the power spectra obtained by FFT were converted to a decibel scale and redistributed into eighteen bins (the Bark-scale bins) which are generally considered as representing a match to the general frequency resolution of the human auditory system.

THE MAPPING NETWORKS

We applied the now classical self-organization procedure proposed by Kohonen (Kohonen, 1988 and Kohonen , 1990) with a program (SOM_PAK) created at the Helsinki University of Technology . The procedure maps the high dimensional input vectors into a two dimensional array of neural units.

A brief review of the procedure is as follows:

Let **X** represent the sequence of input vectors in input space. Each node in the two-dimensional features eighteen weighted input ports . Each node receives the same 18-dimensional input **X.** Let **W** designate the set of weight vectors attached to each node.

Upon presentation of a particular input X^k, the "winning"node is identified in a network location **i** , i.e. the node associated with $\mathrm{Min}\{X^k - W_i\}$.

During the learning phase, all the weight vectors in a set locations i* that includes i as well as a set of its spatial neighbors in the two-dimensional network, are modified according to:

$$DWi^* = r \ n(i,i^*)(X^k - Wi^*)$$

r is the leaning rate, which decreases linearly to 0 as training proceeds. **n(i,i*)** is the neighborhood function.with value 1 at **i* = i** and falls off with the distance between **i** and **i*.**

In our networks the nodes were arranged in an 18 by 12 rectangle with a hexagonal topological structure giving six immediate neighbors to each node . The neighborhood function was a 'bubble' function, equal to 1 if the node was within a given neighborhood radius and 0 everywhere else. The nets were trained for 110,000 steps. For the first 10,000 steps the learning rate began at 0.08 and the bubble radius was 14 nodes. For the final 100,000 steps, the learning rate was reduced to 0.02 and the radius was shrunk to 4.

After extensive training over all subject data, one set (subject 9) was chosen as reference, and the acoustic trace was edited manually to place phonectic and articularoy labels to all utterances This reference data was then presented again to the network for the purpose of branding the nodes

activated by each word.with the appropriate labels. At the end, one can give each responding node a phonetic label, using the most frequent label evoked at this node, along with a 6-vector or real values corresponding to the following measurements (in millimeters):
Tongue tip x (TTX)and y position (TTY);
Tongue blade x (TBX) and y position (TBY);
Tongue dorsum x (TDX) and y position (TDY)
Unresponding nodes were left unlabeled.

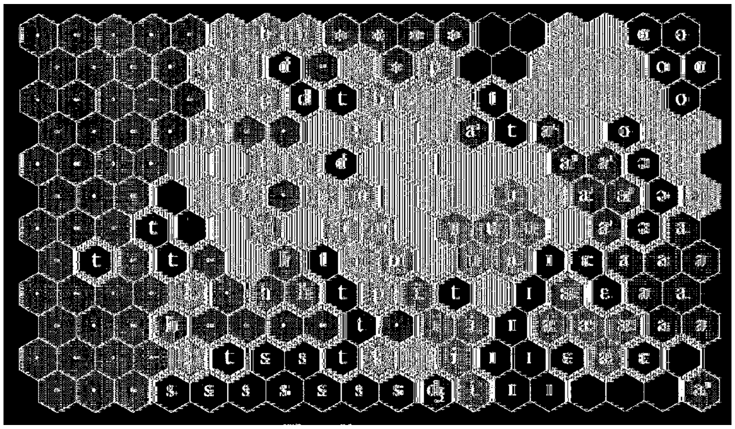

Figure 1:
Phonetic Map

PHONEME ORGANIZATION

Figure 1 shows an example of such phonetic feature map. All phonological symbols used are standard IPA. The asterisks represent a marker tone and the periods represent silence Repeated maps from the same data developed into isomorphic representations regardless of the initial weights.
The maps compared favorably to previous studies on speech using feature maps (Kohonen, 1988). The vowels and consonants clearly separate to different regions on the map. Within the vowels, there are strong groupings for particular sounds and related sounds are adjacent. There is for instance a clear grouping of /æ/,/I/ and /ε/.. There was scattering of groups of /⊃/ amidst the consonants, but in all cases, those input vectors came from the end of an /⊃/ sound on the border of either a consonant or silence. This behavior illustrates the difficulty of labeling transitions between phonemes but it does not affect the articulatory labellings.

Within the consonants, the maps display clear groupings based upon place and manner of articulation. The contiguity of /s/ and /f/; /p/ and /b/; and

/l/ and /Ì/ all are as expected given the similarity of the acoustic signals of these sounds.

SPEECH TRAJECTORIES

During speech, (here a male speaker saying the word 'scar' repeatedly) a dynamic path is traced on the map (figure 2). . The path starts in the /s/ and proceeds to the silence before a burst from the /k/. When the burst occurs, the placement of the /k/ shows much variability, but usually land on silence or another consonant close to the /k/. As the beginning of the vowel begins, there were again fluctuations in the first 2 windows (around 0.03 seconds) before settling in to /a/. As the end of the word draws nearer, the vowel shifts to /a'/ and eventually to /Ì/.

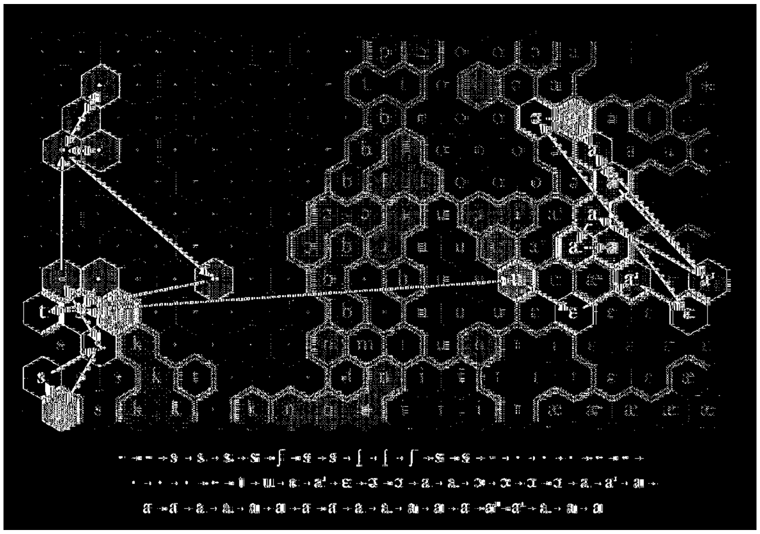

Figure 2:

Trace of the word "*scar*" through the phonetic map

The traces showed good stability except for some difficulty in finding.the labelled /k/s consistently, but often landed on neighboring cells of the /k/.

ARTICULATORY MAPS

The acoustic feature maps are labelled with articulator positions in the same way as with the phonemes. Figure 3 shows the map from figure 2 labeled with TTX position in mm,. The gray scale on the nodes shows groups based upon the phoneme labelling

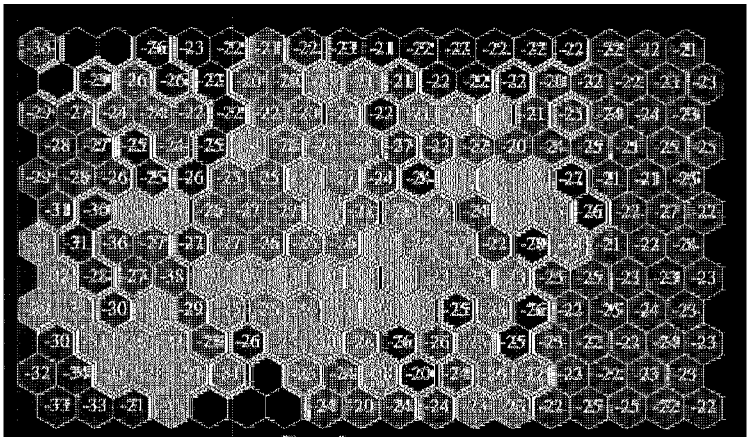

Figure 3:
The map of Figure 1 labeled with Tongue abcissa (TTX)

In general, adjacent nodes respond to similar inputs, however variations in data density will cause some nodes to be closer to than to others after training.

DISCUSSION

There was some scattering of labels, especially noticeable among the vowels, usually caused by the consonant-like transition from the edges of the vowels.
If the labels were not phonemes, but smaller units made up of consonant-vowel transitions and vowel steady-states, those border areas would become separate groups.

CONTINUATION WORK

Three areas of future work can be identified:
1) Traces though the maps when speech is uttered, as had been done for maps labelled with phonemes, should be done on maps labelled with articulations. Initial investigation done on the first map trained on bark bins shows that although there is a large discontinuity in some vowels, the trace for a particular word that uses that vowel remains on one side of the discontinuity. Further investigation of this is called for.
2) Much larger feature maps should be trained on speech data that is better labeled, eliminating or minimizing the overlapping zones that occurred in these maps. This should allow better clustering and more accurate traces through the map.

3) A method for using feature maps to match large dimensional data to large dimensional data through 2 or a similarly small number of dimensions. The authors are actively pursuing

CONCLUSION

This experiment uncovers interesting features of acoustic to articulatory mappings, the use of feature maps This method is powerful enough to analyze other factors involved in acoustic to articulatory mappings. As well, some methods for better using feature maps for speech data were determined.

While there is as yet no evidence that consonants exhibit the same degree of prototype behavior, the possibility of success with respect to modeling vowel perception

ACKNOWLEDGEMENTS

Work was performed at the Los Alamos National Laboratory under contract W-7405-ENG-46 with the University of California.

REFERENCES

Abbs, J.H., Nadler, R.D., and Fujimura, O. X-ray Microbeams Track the Shape of Speech. SOMA: Eng. Human Body, 2. (1988).

Atal, B., Determination of the Vocal Tract Shape Directly From the Speech Wave. J. Acoust. Soc. Am. Suppl. 1 47. (1970).

Grieser, D., and Kuhl, P., Categorization of Speech by Infants: Support for Speech-Sound Prototypes. Develepmental Psychology, 21. (1989).

Hermasky, H., Perceptual Linear Predictive (PLP) Analysis of Speech J. Acoust. Soc. Am., 87. (1990).

Hertz, J., Krogh, A., Palmer, R., Introduction to the Thoery of Neural Computation. Addison-Wesley, Reading MA. (1991).

Kirirani, S., Itoh, K., and Fujimura, O., Tongue-pellet Tracking by a Computer Controlled X-ray Microbeam System. J. Acoust. Soc. Am., 57. (1975).

Kohonen, T., The 'Neural' Phonetic Typewriter. Computer, 21 (3). (1988).

Kohonen, T., The Self-Organizing Map. Proceedings of the IEEE, 78 (9). (1990).

Kohonen, T., Kangas, J., and Laaksonen, J., SOM-PAK, the Self-Organizing Map Program Package. available for anonymous ftp user at the Internet site cochlea.hut.fi.

Kuhl, P., Human Adults and Human infants Show a "Perceptual Magnet Effect" for the Prototypes of Speech Categories, Monkeys Do Not. Perception and Psychophysics, 50 (2). (1991).

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B., Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. Science, 225. (1992).

Ladefoged, P., A Course in Phonetics. Harcourt Brace Jovanovich, San Diego. (1982).

Lieberman, P., and Blumstein, S., Speech Physiology, Speech Perception, and Acoustic Phontetics. Cambridge U.P., Cambridge. (1988).

Nadler, R. D., Abbs, J. H., and Fujimura, O., Speech Movement Research Using the New X-ray Microbeam System. Proceedings of the XIth International Congress of Phonetic Sciences, 1. (1987).

Nittrouer, S., Munhall, K., Kelso, J.A.S., Tuller, B., and Harris, K., Patterns of Interarticulator Phasing and Their Relation to Linguistic Structure. Unpublished manuscript, Haskins Laboratories, 270 Crown St., New Haven, CT. 06511. (1988).

Papcun, G., Hochberg, J., Thomas, T., Laroche, F., Zacks, J., and Levy, S., Inferring Articulation and Recognizing Gestures from Acoustics with a Neural Network trained on X-ray Microbeam Data. J. Acoust. Soc. Am., 92 (2). (1992).

Press, W., Flanner, B., Teukolsky, S., and Vetterling, W., Numerical Recipies in C: The Art of Scientific Computing, Cambridge U.P., Cambridge. (1988).

Protter, M..H., Can One Hear the Shape of a Drum? Revisited. SIAM Rev., 29. (1987).

Wakita, H., Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Waveforms. IEEE Trans. Audio Electro-acoust., 21. (1973).

Wakita, H., Estimation of Vocal Tract Shapes From Acoustical Analysis From the Speech Speech Wave: The State of the Art. IEEE Trans. Acoust. Speech Signal Process. ASSP-27. (1979).

Westbury, J. R., The Significance and Measurement of Head Position During Speech Production Experiments Using the X-ray Microbeam System. J. Acoust. Soc. Am. 89. (1991).