

# Data-driven, non-linear, formant-to-acoustic mapping for ASR

Philip J.B. Jackson, Boon-Hooi Lo and Martin J. Russell

Electron. Elec. & Comp. Eng., Univ. of Birmingham, Birmingham B15 2TT, UK. [p.jackson@bham.ac.uk]

## Abstract

With a view to using an articulatory representation in automatic recognition of conversational speech, two non-linear methods for mapping from formants to short-term spectra were investigated: multi-layered perceptrons (MLPs), and radial basis function (RBF) networks. Five schemes for dividing the TIMIT data according to their phone class were tested. The rms error of the RBF networks was 10 % less than the MLPs', and the scheme based on discrete articulatory regions gave the greatest improvements over a single network.<sup>1</sup>

## 1 Introduction

Using greater training databases and computing power, adjustments have incrementally improved automatic speech recognition (ASR) systems based on hidden Markov models (HMMs) over the past twenty years. Yet the obvious mismatch, between HMM assumptions and acoustic properties of spoken utterances, has prompted research to address their shortcomings: speech patterns are assumed to be piecewise constant, transitions instantaneous, and consecutive observations to be statistically independent. In contrast, *trajectory-based segment models* attempt to capture with an HMM state the dynamics in a sequence of acoustic features, or *segment* [1], as the noisy realization of an underlying trend  $f_a(t)$ , whose parameters  $a$  are typically fixed for any given state  $s_i$ . Despite intuitive appeal, their performance benefits have proven hard to achieve, but have been demonstrated for phone classification [2].

However, trajectories in the acoustic-feature space model articulatory motion indirectly, which manifests itself as movement across, rather than within, frequency bands. Hence, with an articulatory (or pseudo-articulatory) representation, an ASR system can deal directly with coarticulation and speaking rate effects, which is supported by Holmes et al.'s results using vocal-tract resonance frequencies, or formants [3]. Articulatory inversion from acoustic features has non-unique solutions, though, and is prone to error, so performance was only improved by including a confidence measure with the formants. However, previous experience in speech pattern processing suggests that formant identification should emerge as a consequence of recognition, and should be deferred until all relevant

---

<sup>1</sup>The Balthasar project is supported by EPSRC (grant ref. M87146).

available knowledge has been applied. Thus, formant parameters should be embedded in an integrated generative model, and mapped onto a suitable acoustic space for each hypothesised interpretation of the observations.

The mapping from formants to spectral amplitudes is not linear and, to avoid imposing our prejudices, knowledge embodied in training data must be used to learn the most effective transformation. Artificial neural networks offer such properties, with potential for estimating segment trajectories as embedded training [4]. Here, we consider two forms: the multi-layered perceptron (MLP), and the radial basis function (RBF) network. The state segment probability [2], with articulatory-acoustic mapping  $\Theta(\cdot)$ , is

$$b_i(p, q) = b_{i,a}(p, q) = \prod_{t=p}^q \mathcal{N}(y_t; \Theta(f_a(t)), R_y), \quad (1)$$

where  $\mathcal{N}(y_t; f_a(t), R_y)$  denotes a multivariate Gaussian with mean  $\Theta(f_a(t))$  and covariance  $R_y$ , evaluated at  $y_t$ . While the rôle of articulatory representations in human speech recognition remains controversial, we believe that the constraints of a continuous dynamical system (in pseudo-articulatory space), whose parameters can be learnt from speech data, are sufficient for ASR.

## 2 Method

Acoustic features and formants were extracted (10ms frame rate) from the TIMIT database using the Holmes formant tracker [6], as short-time log spectra (32 bands in the range 0–4kHz, in 0.25 dB steps), and F1, F2 and F3 (25 Hz resolution), respectively. Only male speakers’ data were used, split into training, evaluation and test sets. Using spectra enabled more ready interpretation of mapping errors, yet our findings are as applicable to MFCCs.

Speech sounds are traditionally categorized by manner (and place) of articulation. Though similar in outcome, it is more relevant to classify them here according to the discrete combinations of acoustic-source type and vocal-tract configuration. The English phonetic repertory, limited also by physical constraints, has ten discrete articulatory regions (including silence) that depend on whether phonation (voiced/voiceless), frication and/or plosion are present, the nasal cavity is coupled (velum open/closed), and the extent of oral occlusion (open/constricted/closed). With one network per class, a series of networks was obtained for each phone classification scheme (number used): A. speech vs. silence/non-speech (2); B. individual phones (63); C. as in Deng and Ma [5] (10); D. conventional phonetic categories (6); E. discrete articulatory regions (10). A single network was also trained on the whole training set. The proportion of the total error contributed by each of the various phone classes was analysed.

## 3 Multi-layered perceptrons

The MLPs each contained one hidden layer with sigmoidal activations, and a linear output layer, had 3 inputs and 32 outputs (normalised to range from 0 to 1), and were given a small, random initialisation. The number of

hidden units (4–100) had little effect, except on the rate of convergence, contrary to what had been expected. In experiments for a single speaker, MLP performance was improved by adjusting learning and momentum terms for gradient descent, and the number of training iterations. Similar adjustments gave no significant advantage, when applied to all male speakers, for whom training converged within forty iterations (conjugate gradients behaved comparably). Results with 10-hidden-unit networks, after batch gradient-descent training (40 its), are presented in Figure 1 (upper), for the single network and schemes: A. 0 silence, 1 speech; C. 0 silence, 1 vowel, 2 glide/liquid, 3 nasal, 4 voiceless fricative, 5 /s,ch/, 6 voiced fricative, 7 /z,jh/, 8 voiceless stop, 9 voiced stop; D. 0 stop/silence, 1 vowel, 2 glide/liquid, 3 nasal, 4 fricative, 5 affricate; E. 0 silence, 1 vowel/glide, 2 liquid, 3 nasal, 4 voiced stop closure, 5 voiceless stop closure, 6 voiced plosive, 7 voiceless plosive, 8 voiced fricative, 9 voiceless fricative.

Occupying half the speech frames with high-amplitude spectra, vowels contribute 60% of the initial rms error at the outputs. They are well-modelled by an MLP trained on speech frames, and the error reduced by 90% when trained exclusively on vowel data. After training, the residual error is highest for plosive consonants, since their formant estimates are highly inaccurate. With a single network, the smallest reduction in network error is seen for silence frames, for which formants are clearly meaningless, hence marking them as a separate class derived substantial benefit: overall error was reduced to 2.5 dB with scheme B (cf. 4.2 dB, single MLP).

## 4 Radial basis function networks

The RBF networks had two layers too: the first responded to the formant inputs via non-normalised Gaussians, the second linearly combined these responses to yield the estimated output for each spectral bin. In training, RBF centroids were determined by  $k$ -means clustering (random initialisation), then the linear layer by singular value decomposition [7].<sup>2</sup> Unlike the MLP, the number of hidden RBF units strongly affected the evaluation-set error, which fell as number increased. Accordingly, 20 centroids were used, and the results are shown in Figure 1 (lower), demonstrating universal improvement compared to the MLPs, although the error pattern is almost identical across phone classes, and the five schemes ranked equally. Nevertheless, the effect of class-specific networks was greater for the better-performing RBFs, and scheme E is just one point from phone-specific scheme B, 2.3 dB vs. 2.2 dB.

## 5 Discussion

Figure 2 illustrates differences in performance of the two kinds of mapping with an example speech file. Formants do influence the spectral shape from the single MLP (e.g., frames 120–150), but the output tends to be blurred spectrally, rather static, and of average amplitude throughout. The MLP fails to match silences (frames 0–12, and c.80), and much spectral detail is lost. The RBF network, however, reproduces the spectra more faithfully, but

---

<sup>2</sup>Removing singular values below machine precision avoided any singularities in the pseudo-inverse.

does not model gross changes in amplitude. Both outputs are predominantly grey because there is no amplitude input, nor any information of source or vocal-tract configuration, and so training is driven by outlying data.

Formants, being predominantly local features in a spectrum, were better modelled by the RBFs than the MLP sigmoids, whose influence is less localised. Moreover, training of the MLP's first and second layers was unbalanced, and quickly found a local optimum dominated by the second linear layer. This problem could not be offset by careful scaling of the data or initialisation. The two-stage RBF optimisation was much more robust.

In tests, RBF networks consistently mapped better than MLP networks by  $\sim 0.4$  dB, or 10% of the residual rms error. The classification schemes' effect was greater, reducing the error to 2.3 dB. Patterns of error across the different phonetic categories were similar for single networks and class-specific networks. Generally, single networks performed worst, next the speech/non-speech classification (scheme A), then the phone-class schemes (C–E), and finally the phone-specific networks yielded the smallest errors (B). The best phone classification, scheme E, was based on discrete regions of the articulatory space, benefitting the RBFs more than for the poorly-performing MLPs. These results clearly demonstrate that spectral information expressed as formants can be represented by a two-layer neural network, appropriately trained. Extensions of this research include adding inputs and embedding the network within a phoneme recognition system; see <http://web.bham.ac.uk/p.jackson/balthasar/>.

## References

- [1] Ostendorf, M., Digalakis, V., and Kimball, O.A., "From HMMs to segment models: a unified view of stochastic models for speech recognition", *IEEE Trans. SAP*, 4(5): 360–378, 1996.
- [2] Wendy J. Holmes, and Martin J. Russell, "Probabilistic-trajectory segmental HMMs", *Computer Speech and Language*, 13(1): 3–37, 1999.
- [3] Holmes, W.J., Holmes, J.N., and Garner, P.N., "Using formant frequencies in speech recognition", *Proc. Eurospeech'97*, 2083–2086, 1997.
- [4] Richards, H.B., and Bridle, J.S., "The HDM: A segmental hidden dynamic model of coarticulation", *Proc. IEEE-ICASSP'99*, 357–360, 1999.
- [5] Deng, L., and Ma, J., "Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics", *J. Acoust. Soc. Am.*, 108(6): 3036–3048, 2000.
- [6] John N.J. Holmes, "Speech processing system using formant analysis", US patent 6292775, Sept. 2001.
- [7] Bishop, C., "Neural Networks for Pattern Recognition", Clarendon Press, Oxford, UK, 1995.

## List of Figures

1	Test results for MLPs (top) and RBF networks (bottom), using a single network (lighter grey bar, overall; + dashed, per class), and phone-class schemes A to E (darker grey bar, overall; o solid, per class). . . . .	6
2	Single network outputs for evaluation example <code>train/dr1/mcpm0/si1194.wav</code> : (from top) original spectrogram, MLP, and RBF network. . . . .	7

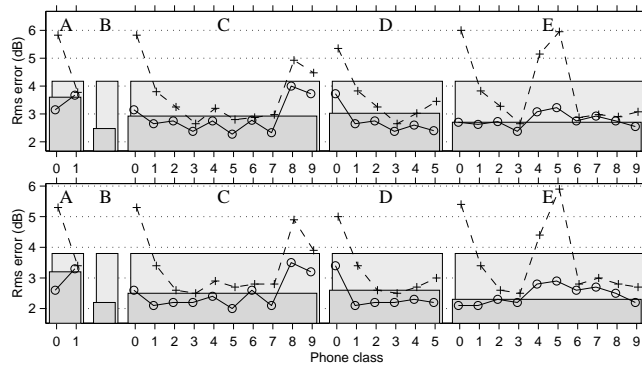


Figure 1: Test results for MLPs (top) and RBF networks (bottom), using a single network (lighter grey bar, overall; + dashed, per class), and phone-class schemes A to E (darker grey bar, overall; o solid, per class).

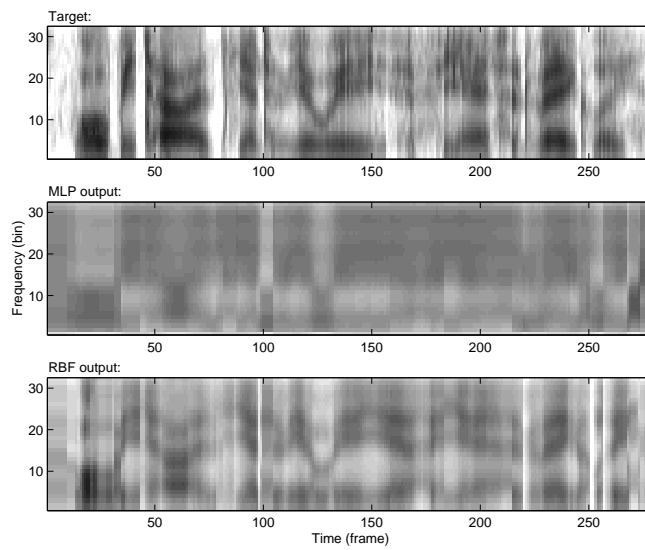


Figure 2: Single network outputs for evaluation example `train/dr1/mcpm0/si1194.wav`: (from top) original spectrogram, MLP, and RBF network.