S3.9

THE META-PI NETWORK: CONNECTIONIST RAPID ADAPTATION FOR HIGH-PERFORMANCE MULTI-SPEAKER PHONEME RECOGNITION

John B. Hampshire II Alex H. Waibel

School of Computer Science, Carnegie Mellon University
Pittsburgh, Pennsylvania 15213-3890, USA

ABSTRACT

We present a multi-network Time-Delay Neural Network (TDNN)-based connectionist architecture that allows us to perform multi-speaker phoneme discrimination (/b,d,g/) at the speaker-*dependent* recognition rate of 98.4%. The overall network gates the phonemic decisions of modules trained on individual speakers to form its over-all classification decision. By dynamically adapting to the input speech and focusing on a combination of speaker-specific modules, the network outperforms a single TDNN trained on the speech of all six speakers (95.9%). To train this network we have developed a new form of multiplicative connection that we call the "Meta-Pi" connection. We illustrate how the Meta-Pi paradigm implements a dynamically adaptive Bayesian MAP classifier. It learns — without supervision — to recognize the speech of one particular speaker (99.8%) using a dynamic combination of internal models of *other* speakers exclusively. The Meta-Pi model is a viable basis for a connectionist speech recognition system that can rapidly adapt to new speakers and varying speaker dialects.

I. INTRODUCTION

The objective of this research was to extend the speaker dependent phoneme recognition performance of TDNNs [1] to a multi-speaker task, focusing on ways to build a connectionist phoneme classifier that could utilize an integrating superstructure to combine a number of speaker-dependent sub-networks (or "modules") into a larger structure. We were motivated in this direction by two factors:

- Issues of modularity and scaling [3,4] (i.e., the need for computationally efficient modular connectionist structures that could form the basic building blocks of much larger integrated systems).
- The tradeoff in speech recognition systems that strive for speaker independence by training a single large speech model with speech from a broad collection of speakers: speaker-independent recognition rates are usually well below speaker-*dependent* recognition rates. As an example, our earlier experiments using a single TDNN for the six-speaker /b,d,g/ task yielded a recognition rate of 95.9% [4] — well below the 98.5% result for the single-speaker /b,d,g/ task.

We wanted our speech model to have the flexibility necessary for robust recognition across a range of speakers with widely varying vocal tract characteristics, yet we wanted speaker dependent recognition rates; our notion of flexibility was based on how readily the classifier could adapt to a novel speaker. Psychoacoustic studies have suggested that people actively adapt to new speakers within a few uttered syllables [5]. Our goal was to develop a modular connectionist architecture that could adapt to new speakers, thereby focusing the recognition process on an internal model of a single speaker or a *set of relevant speakers* This led to two basic levels of adaptation:

1. **Rapid Adaptation**: The network should recognize a novel speaker without additional training, using an adaptive combination of internal speaker-dependent modules to model the new speech.
2. **Slow Adaptation**: If rapid adaptation were still inadequate for modeling the new speaker, the network should initiate a learning process to develop and "fine tune" a suitable additional model for properly representing the new speaker.

We describe the "Meta-Pi" network, a modular connectionist structure that addresses issues of rapid adaptation and forms a viable basis for more complex adaptive systems.

II. SPEECH DATA

The nature of the speech data used in this experiment is detailed in [1]. We obtained isolated Japanese word utterances from six professional announcers (2 female, 4 male). From this phonetically balanced database we then obtained approximately 200 training and 200 testing tokens for each of the three voiced-stops /b,d,g/ from each speaker. This data formed six training and six testing /b,d,g/ data sets: one for each of the six speakers. We also mixed all six training sets to form a "global" training set, and repeated the process to build a disjoint global testing data set.

III. THE META-PI PARADIGM

The general form of the Meta-Pi network is illustrated in figure 1. It actually comprises many TDNN modules — all working in parallel to classify the input speech signal. Each of these modules is trained to classify the speech of one particular speaker. The N global network outputs $O_1, \ldots, O_N$ correspond to N possible phonemes to be detected in the input speech signal. The speaker-dependent outputs $\rho_{k,1}, \ldots, \rho_{k,N}$ correspond to their global counterparts $O_1, \ldots, O_N$. In the case of the /b,d,g/ task trained with six speakers $K = 6$ and $N = 3$ (for the three phonemes). The most active output of the global structure constitutes the structure's classification of the input token[1].

The N connections linking the N outputs $\{\rho_{k,1}, \rho_{k,2}, \ldots, \rho_{k,N}\}$ of each of the K speaker-dependent modules to the N global outputs $\{O_1, O_2, \ldots, O_N\}$ via their respective Meta-Pi unit M_{π_k} are shown as single arrows. The network's nth global output is given by

$$O_n = \frac{1}{\mu} \sum_K \rho_{k,n} \cdot M_{\pi_k} \qquad (1)$$

where

$$\mu \triangleq \sum_K M_{\pi_k} \qquad (2)$$

[1]In order to keep figure 1 compact and visually clean, the output nodes of each module and the global outputs have been aligned vertically.

CH2847-2/90/0000-0165 $1.00 © 1990 IEEE

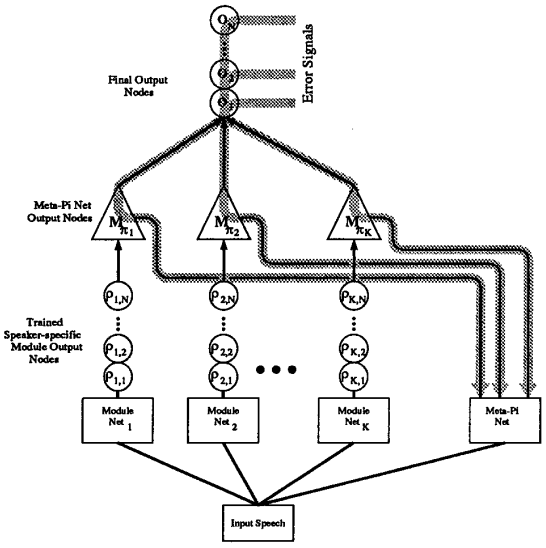

Figure 1: The Meta-Pi network: K speaker-dependent modules linked by an integrating superstructure.

$$\begin{aligned} \mu &> 0 \\ 0 \leq M_{\pi_k} &\leq 1 \quad \forall k \end{aligned}$$

The gray arrows in the figure illustrate how the error signal created by network's output during learning backpropagates [6] through the Meta-Pi integrating superstructure. Detailed equations describing error gradient computations associated with the Meta-Pi network are contained in [7]. All of these computations hinge on the partial derivatives relating the network's global outputs to the speaker-dependent module outputs and the Meta-Pi integrating superstructure's outputs:

$$\frac{\partial O_n}{\partial M_{\pi_k}} = \frac{1}{\mu}\left[\rho_{k,n} - O_n\right] \tag{3}$$

IV. A BAYESIAN VIEW OF META-PI

Let us consider for a moment that each global output O_n of the Meta-Pi structure represents an estimate of the continuous-valued probability of obtaining the input speech signal $\overline{A}$ when $\overline{A}$ is truly meant to represent the phoneme[2] D_n:

$$O_n = \left\langle P(\overline{A}|D_n)\right\rangle \tag{4}$$

In reality, the probability $P(\overline{A}|D_n)$ given in (4) is conditional in nature — principally affected by the dialectal characteristics, vocal tract properties, physical and emotional states, etc. of the speaker actually uttering the input signal $\overline{A}$. We bind all of these probabilistic conditions into the state variable $\overline{S}$ (denoting speaker) so that (4) is more precisely expressed as

$$O_n = \left\langle P(\overline{A}|D_n, \overline{S})\right\rangle \tag{5}$$

Assuming the global outputs $O_1, \ldots, O_N$ represent all *possible* phonemes given $\overline{A}$, one can express the probability that the input $\overline{A}$ represents the phoneme D_n by Bayes rule:

$$P(D_n, \overline{S}|\overline{A}) = \frac{P(\overline{A}|D_n, \overline{S}) \cdot P(D_n, \overline{S})}{\sum_{j=1}^{N}\left[P(\overline{A}|D_j, \overline{S}) \cdot P(D_j, \overline{S})\right]} \tag{6}$$

If the joint prior probabilities $P(D_j, \overline{S})$ are equivalent for all j, (6) reduces to

$$P(D_n|\overline{A}) = \frac{P(\overline{A}|D_n, \overline{S})}{\sum_{j=1}^{N} P(\overline{A}|D_j, \overline{S})} \cong \frac{O_n}{\sum_{j=1}^{N} O_j} \tag{7}$$

Clearly, (7) yields optimal classification performance when O_n constitutes an accurate estimate of $P(\overline{A}|D_n, \overline{S})$ for all n. If we consider $\rho_{k,n}$ (the nth output of the kth speaker-dependent module) in the same probabilistic light that we view O_n, then

$$\rho_{k,n} = \left\langle P(\overline{A}|D_n, \overline{S}_k)\right\rangle \tag{8}$$

Using (1), (7), and (8) we can form the following relationships[7]:

$$\begin{aligned} \frac{O_n}{\sum_{j=1}^{N} O_j} &= \frac{\sum_K \rho_{k,n} \cdot \frac{M_{\pi_k}}{\mu}}{\sum_{j=1}^{N}\sum_K \rho_{k,j} \cdot \frac{M_{\pi_k}}{\mu}} \\ &= \frac{\sum_K \left\langle \rho(\overline{A}|D_n, \overline{S}_k)\right\rangle \cdot \left\langle P(\overline{S}_k|\overline{A})\right\rangle}{\sum_{j=1}^{N}\sum_K \left\langle \rho(\overline{A}|D_j, \overline{S}_k)\right\rangle \cdot \left\langle P(\overline{S}_k|\overline{A})\right\rangle} \end{aligned} \tag{9}$$

Thus, if one views the Meta-Pi network's outputs as representing the conditional probability that a particular speaker-dependent model applies to the given input utterance $\overline{A}$ and if the set of speaker-dependent models $\{ S_1, \ldots, S_K \}$ is all-inclusive for $\overline{A}$, (9) forms the following approximation on the basis of linear superposition:

$$P(\overline{A}|D_n, \overline{S}) \cong \sum_K P(\overline{A}|D_n, \overline{S}_k) \cdot P(\overline{S}_k|\overline{A}) \tag{10}$$

where

$$\left\langle P(\overline{S}_k|\overline{A})\right\rangle = \frac{M_{\pi_k}}{\mu} \tag{11}$$

In the Bayesian context, the Meta-Pi network maximizes the a posteriori probability of a correct global phoneme classification **by learning to compute the conditional prior probabilities** $P(\overline{S}_k|\overline{A})$ $\forall k$ **dynamically**. It is critical to note that the Meta-Pi network learns to compute these speaker-specific conditional priors without any explicit knowledge of speaker identity. The Meta-Pi integrating superstructure's weights are adjusted solely on the basis of how well the global network performs the global classification task.

V. THE META-PI NETWORK'S PERFORMANCE

The Meta-Pi network learns early during the training phase that gender plays a critical role in accurate phoneme recognition. As a result the Meta-Pi network learns — without supervision — to group speakers by gender. Figure 2 illustrates this phenomenon. The figure shows the unique connections between the input layer of the Meta-Pi TDNN and its first hidden layer [3]. Twelve groups of connections depict the weights linking three 16-coefficient spectra (48 units) of the input layer to each of twelve first-hidden-layer units. Positive connections are white, negative connections are black, and the magnitude of each connection is proportional to the size of its corresponding rectangle. It is clear from figure 3 that the input to first-hidden-layer

[2]Hinton proffers a very similar interpretation of Multi-layer Perceptron outputs as his rationale for the Cross Entropy (CE) objective function [8].

[3]See [1] for details of the TDNN's temporally constrained connection topology.

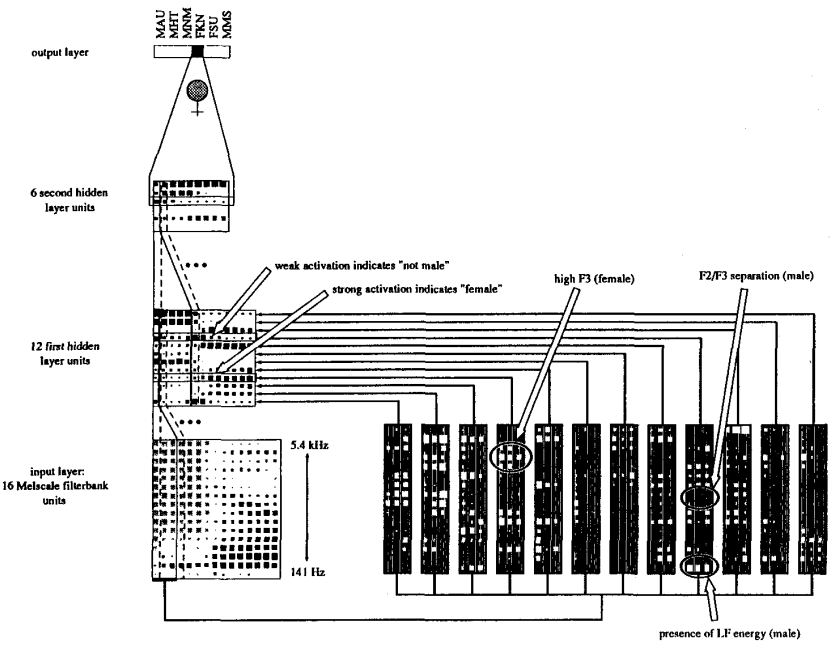

Figure 2: The Meta-Pi network's input-to-hidden-layer connections use gross formant features to implement the Meta-Pi integrating function.

connections are block-like (i.e., positive and negative connections to each first-hidden-layer node are clustered in very regular blocks). These blocks tend to correspond to formant locations for the various speakers. The figure illustrates two formant features (F2/F3 separation and the presence of low-frequency [LF] energy) that the Meta-Pi network has learned to use for detecting male speech. Likewise, the Meta-Pi network has learned that a relatively high-frequency third formant (F3) indicates a female's speech. The Meta-Pi network has learned to rely heavily on formant characteristics in order to maximize its global recognition performance.

Although the Meta-Pi network is capable of specific speaker identification [7] the percentage of utterances for which it specifically identifies a single module for the global recognition task is actually low — less than 30%. Figure 3 illustrates a much more typical mode of operation for the Meta-Pi network. In this figure the speech token is actually uttered by male MHT, but the Meta-Pi network associates the input stimulus with three male modules and produces an unambiguous correct global recognition result. The network has learned to perform explicit speaker identification when the input signal possesses features that are unique to a particular individual, yet it has also learned that many utterances are prototypical of a *group* of speakers (e.g., males or females). In cases where the utterance is prototypical the Meta-Pi network attributes the speech to a collection of speakers associated with the prototype.

Another notable aspect of figure 3 is that the Meta-Pi network does not use the MHT module to recognize this utterance. *In fact, when fully trained, the Meta-Pi superstructure* **never** *uses the MHT module to classify speech tokens from* **any** *of the six speakers.* Indeed the superstructure learns to model speaker MHT with a dynamic combination of *other* male speakers and, in so doing, still achieves a 99.8% recognition rate on the speech of MHT. An detailed analysis of this phenomenon is given in [7]. In short, the Meta-Pi superstructure learns that it can accurately model MHT's speech using combinations of other male modules. Additionally, it finds that there is little utility in using the MHT module to recognize speech tokens from any of the other speakers . While its recognition rate is high for $1/6$ of the training data (MHT's speech), it is quite low for the remaining $5/6$ of the training data. As a result, the Meta-Pi superstructure learns that there is no marginal utility in using the MHT module. In effect, the Meta-Pi superstructure removes the MHT module from its "module database".

These results hold promise for connectionist pattern recognition systems for two reasons:

- First, the Meta-Pi integrating superstructure clearly develops general and flexible models of conditional stimuli (i.e., the characterization of basic speaker types based on very gross formant features [figure 2]); these general models appear to have significant potential for processing novel speakers without any modification of the Meta-Pi combinational superstructure (rapid adaptation). The Meta-Pi superstructure's ability to learn how to recognize a particular speaker using a dynamic combination of models trained on *other* speakers indicates that the architecture can represent novel (unknown) stimuli using existing internal models of known stimuli. This, in turn, suggests that the architecture can *adapt* to novel stimuli, using dynamically altered combinations of known stimuli. Because the integrating superstructure is relatively simple, one would expect this adaptation process to be rapid. We are investigating this issue at present.

- Second, the integrating superstructure demonstrates its ability to maintain its own database of relevant stimuli *without any external supervision beyond the global phoneme recognition objective*: in fact the Meta-Pi superstructure learned that its model for the MHT stimulus was unnecessary for robust performance of the global phoneme recognition task,

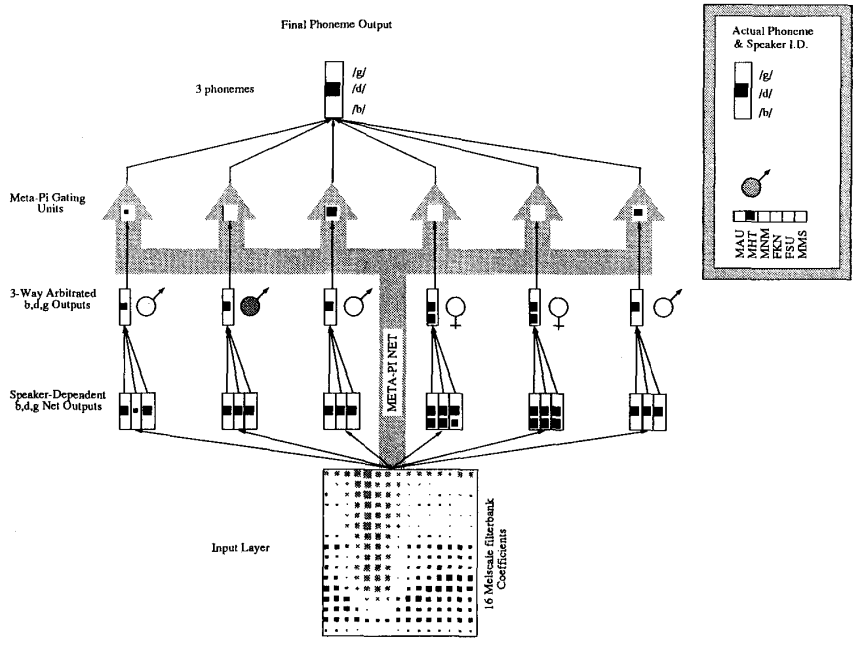

Figure 3: The Meta-Pi superstructure recognizes the speech of male MHT using a dynamic combination of *other* male speaker modules.

and removed the MHT module from the modular structure. This "behavior" could form a basic element of autonomous connectionist speech recognition systems. Such systems would develop and maintain their own database of speaker models, adapting to new speakers when possible and spawning new speaker-dependent learning processes when necessary. These systems could also eliminate redundant or obsolete speaker-specific modules when appropriate.

VI. CONCLUSION

In this report we have presented the Meta-Pi network, a connectionist architecture for multi-speaker phoneme recognition. The Meta-Pi network achieves a 98.4% recognition rate on the voiced stops /b,d,g/ of six speakers (2 females, 4 males). This recognition rate constitutes a significant improvement over the 95.9% recognition rate we obtain from a single TDNN trained on speech from all six speakers. We have described how the Meta-Pi network constitutes a Bayesian MAP connectionist classifier with potential for adaptive speech recognition systems.

REFERENCES

[1] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., *Phoneme Recognition Using Time-Delay Neural Networks,* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-37, No. 3, March, 1989, pp. 328-339.

[2] Lang, K. "A Time-Delay Neural Network Architecture for Speech Recognition," Ph.D. Dissertation, *Carnegie Mellon University technical report CMU-CS-89-185*, July, 31, 1989.

[3] Waibel, A., Sawai, H., Shikano, K., *Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks,* Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing, May, 1989, pp. 112-115.

[4] Hampshire, J., Waibel, A., *Connectionist Architectures for Multi-Speaker Phoneme Recognition,* Carnegie Mellon University Technical Report CMU-CS-89-167, August, 1989.

[5] Kato, K., and Kakehi,K., "Individual Speaker Differences in Monosyllabic Speech Perception," NTT Technical Report, 1986. Also published in the *Journal of the Acoustical Society of Japan*, 1987.

[6] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., "Learning Representations by Backpropagation Errors," *Nature*, vol. 323, pp. 533-536, October, 1986.

[7] Hampshire, J., Waibel, A., *The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Pattern Recognition,* Carnegie Mellon University Technical Report CMU-CS-89-166, August, 1989.

[8] Hinton, G. E., "Connectionist Learning Procedures," *Carnegie Mellon University Technical Report CMU-CS-87-115 (version 2)*, December, 1987, pg. 14.

*The authors gratefully acknowledge ATR Interpreting Telephony Research Laboratories, the National Science Foundation, and Bell Communications Research for supporting this research.