

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 19 (1993-1994)
Indiana University

Models of Speaker-Dependent Speech Recognition¹

Michael L. Kalish and Lynne C. Nygaard

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹This research was supported by NIH Research Grant DC-00111 and PHST Training Grant MH-19879-02 to Indiana University in Bloomington, IN.

Abstract

We first propose a list of possible models of the interaction of lexical and indexical information and processing. A cursory review of the extant psychological data allows rejection of all but one model class; that in which lexical and indexical information and processing are mixed but not synonymous. One such model, which we term generalization, is developed and tested against human learning data. This model is rejected, because it requires forgetting of old familiar speakers in order to explain improvements for newly learned speakers. A second model, which we call true transfer, is proposed with two conceptualizations. Under either description, the model predicts equivalences in the role of long-term knowledge of speaker identity and short-term adaptation to unfamiliar speech.

Models of Speaker-Dependent Speech Recognition

Speaker identity and linguistic content are two different perceptual objects; they come from the same physical source and are carried by the same medium. The question is thus raised as to their relationship both in the signal and in perception. Four possible relationships exist. Either (a) the information is separate in the signal (i.e., carried by independent dimensions) and the perceptual processes occur in isolation, or (b) the information is separate in the signal but processing of identity and linguistic content are related, or (c) the information is overlapping but the processes are separate, or (d) the information and the processes are both overlapping

The historical view has been that (c) is the case. Talker characteristics are thought to add variability to the aspects of the speech signal which are important for recovery of linguistic content. The process of speech perception is thought to involve normalization of the acoustic signal in which speaker information is actively discarded (Halle, 1985; Joos, 1948; Summerfield & Haggard, 1973). Similarly, speaker identification (at least in machines) involves sampling over extended durations in order to obtain text-independent estimates of speaker-dependent acoustic attributes.

Recent experiments, as well as neuropsychological evidence, suggest that this approach is not completely correct. Evidence from a number of studies (Nygaard, Sommers, & Pisoni, 1994; Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992; Remez, Fellowes, Pardo, & Rubin, 1993) has demonstrated that sensitivity to the linguistic content of an utterance is modulated by the familiarity of the talker's voice. These data suggest that (d) is most likely to be true, that information for linguistic and speaker-specific objects of perception are mixed both in the signal and in the head.

Van Lancker, Cummings, Kreiman, and Dobkin (1988) have demonstrated a dissociation between speaker and speech perception which bears directly on the issue of the overlap between indexical and linguistic processing. In phonagnosics, interpretability of speech remains high even though identifiability of speakers is lost. This suggests that, while similar signal sources inform about both perceptual objects, the objects are processed separately at least somewhere in the brain. Thus, the ability to recognize a speaker by their voice is not required in order to perceive their speech. However, many of the acoustic dimensions important for speech recognition are similar or identical to those useful for speaker identification (Remez, et al., 1993).

Listeners are able to use information about speakers to increase the interpretability of their utterances, but their linguistic processing is apparently unimpaired when that speaker-specific information is unable to lead to judgments of speaker identity. The suggestion then is that linguistic processing (LP) and speaker identification (SI) share a common level of processing above the transduction of the signal, but below decision processes.

One model of this communality of processing is based on *generalization*. In generalization, it is assumed that training on SI leads to changes in the perceived similarity of words. That is, novel utterances by a familiar individual are more similar to the listener's lexical representations than are the novel utterances of unfamiliar speakers. This change in turn makes samples of the trained speaker's speech more interpretable, since they are closer to the acoustic 'templates' or long-term memories in the listener's lexical inventory.

In contrast, a *transfer* model supposes that familiarity with a voice (for example, through SI training) leads to a change in processing or representation that is beneficial to performance in both SI and

LP. An analogy can be made to physical exercise, in which strengthening of a muscle can lead to increased performance in two different tasks, not because of increased similarity between the tasks but because the underlying ability to do both tasks has been increased. Following the explication of generalization's failure to account for the effect of speaker familiarity on linguistic processing (Nygaard, et al., 1994), we propose a transfer model in which SI training 'strengthens' the 'mental muscle' upon which the perception of both speaker-specific and linguistic objects rely.

Nygaard, Sommers, and Pisoni (1994)

The results we are modeling come from an experiment conducted by Nygaard, Sommers and Pisoni (1994) that investigated the relationship between the encoding of information about a speaker's voice and the analysis of the linguistic content of a speaker's utterance. By assessing the effects of talker familiarity, or long-term memory for a speaker's voice, on the speech perceptual process, Nygaard et al. (1994) sought to determine the inter-dependence of speaker identification and linguistic processing.

To assess this relationship, listeners were asked to learn to identify the voices of 10 talkers (5 male; 5 female) over a nine day training period. At the end of the training period, the role of talker recognition in the perception of spoken words was evaluated to determine if the ability to identify a talker's voice was independent of phonetic analyses. If learning to identify a talker's voice was found to affect subsequent word recognition performance, the results would argue for a direct link between the mechanisms responsible for the encoding of talker information and those that underlie phonetic perception. Thus, at test, subjects identified the lexical content of words presented in noise produced either by the speakers they encountered in training or by a set of novel unfamiliar talkers.

Insert Figure 1 about here

Figure 1 shows Nygaard et al.'s (1994) basic results. Mean intelligibility of words produced by familiar and unfamiliar talkers is plotted as a function of signal-to-noise ratio. Subjects who heard familiar voices in the word intelligibility task were better at identifying novel words at each signal-to-noise ratio than control subjects who heard unfamiliar voices. Thus, it appears that the ability to explicitly identify a talker's voice improved intelligibility of novel words produced by the same talkers.

Generalization

One account of Nygaard et al.'s (1994) results is that familiarity with a talker's voice increases the similarity between words spoken by that talker and representations of the same words in memory. This account, which is based on strict generalization, suggests that the influence of learning a talker's voice on word intelligibility is a result of generalization from specific talkers and words heard in training to the highly similar words produced by familiar speakers encountered in the word intelligibility transfer task.

Insert Figure 2 about here

The generalization model depicted in Figure 2 makes two fundamental assumptions. First, it is assumed that voice and lexical processing share a common representation or process. This assumption is

Intelligibility of Words in Noise

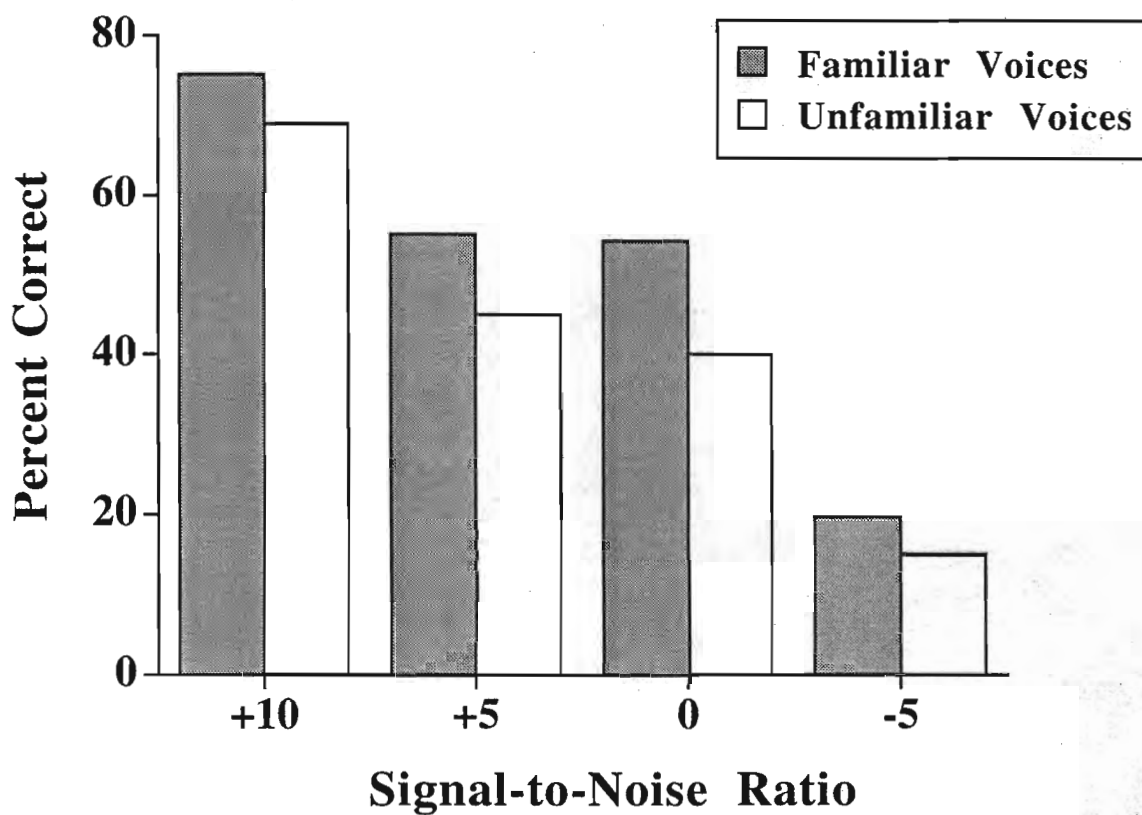


Figure 1. Mean intelligibility of words presented in noise for trained and control subjects. Percent correct word recognition is plotted at each signal-to-noise ratio. (From Nygaard et al., 1994).

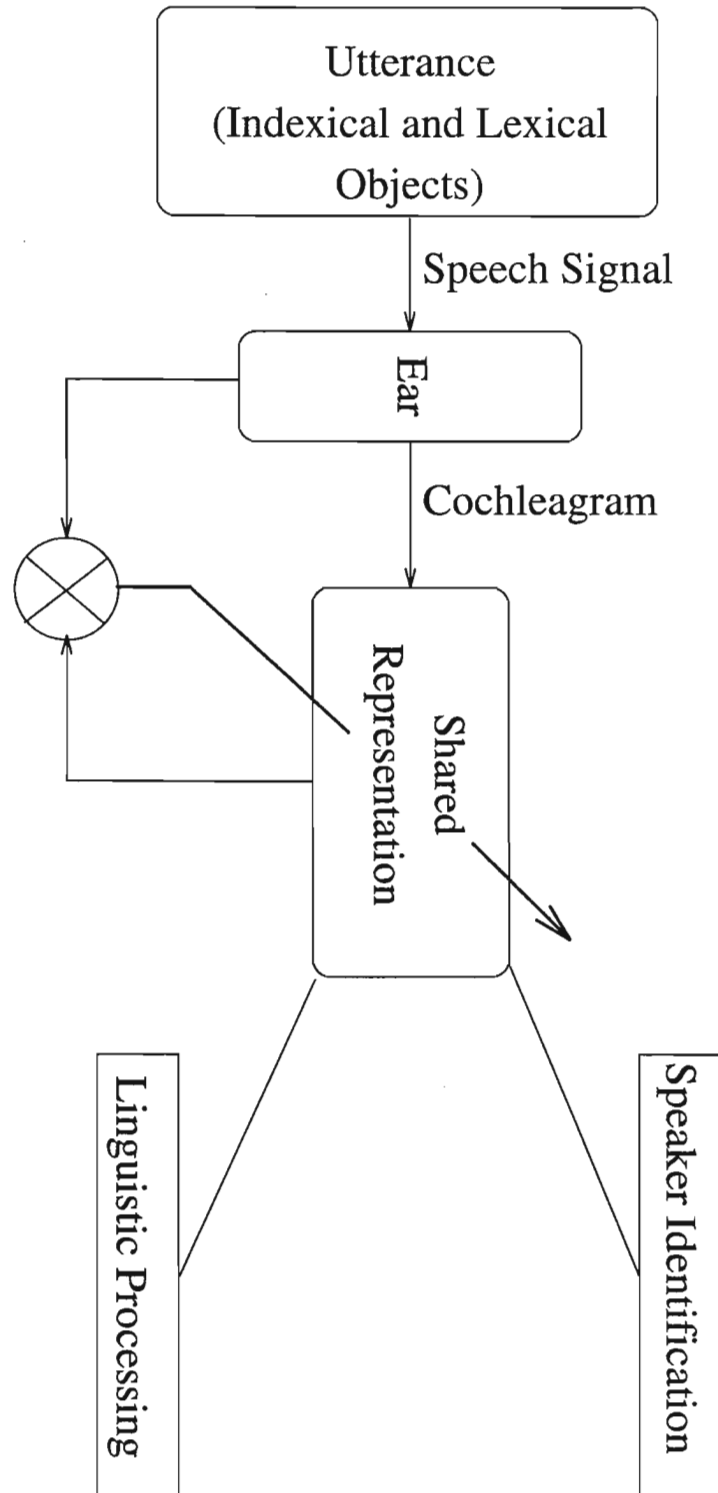


Figure 2. The generalization model. The shared representation alone is modified by experience.

required to account for the effect of differences between familiar and unfamiliar talkers in lexical interpretation. Second, significant change in this shared representation must occur during perceptual learning of novel voices. For familiarity to work through a change in the similarity structure of the representation, the change must be large and widespread.

A network formalization of the generalization model was used to assess the plausibility of the model's assumptions. The network was composed of linear sigmoid units, fully interconnected in a recurrent structure. Both the input and output of the network were vectors representing the state of the auditory nerve as it is modulated by cochlear activation. That is, the network was trained to predict the next slice of a cochleagram (Lyon, 1982). One set of units received the input at time t and one set received training signals of the slice taken at time $t + 1$ (see Figure 3). The entire network comprises the 'shared representation' needed for the generalization model.

Insert Figure 3 about here

Cochleagrams generated by Slaney's (1988) implementation of Lyon's (1982) computational model served as both input and targets for the network. Cochleagrams reflect the filtering characteristics of the peripheral auditory system, and provide a representation that is hypothesized to correspond to auditory nerve activations. This physical description of the speech signal serves to provide the bases for judgments of similarity of tokens of different words spoken by different talkers.

Networks were first trained to predict the cochleagrams of a large ($N=10$) set of speakers, five male and five female. The stimuli were a subset of those used in the original Nygaard et al. (1994) study. Each of the 500 tokens was resampled until the network reached a criterion overall error level. A network was considered pretrained after this phase. Following pretraining, the network then entered a training phase in which four novel speakers (two male and two female) each produced twenty words. The twenty words were the same for all four speakers, and were drawn from the pretraining set of fifty. Finally, during the test phase, network weights were fixed and performance was assessed for voices learned during pretraining (now, 'unfamiliar') and training ('familiar') phases, as well as on words from each set. Performance was evaluated both before and after the training phase.

Consistent with the results of Nygaard et al. (1994), the fully trained networks made fewer errors in predictions of training set words spoken by familiar speakers than had been made following pretraining alone. Performance was also better than for the same words spoken by members of the pre-training set. In addition, words from the pretraining set alone ('novel' words) showed the same pattern of performance, with the fully trained networks showing a superiority for familiar speakers. However, this difference was due as much to a decrement in performance on the original training set as to improvement of performance for the familiar talkers. For human listeners, the un-trained voices were understood just as accurately before and after familiarization with a small set of speakers (Figure 4). This argues that the representations of those words were just as accurate before and after training, which is not true of the network model.

Insert Figure 4 about here

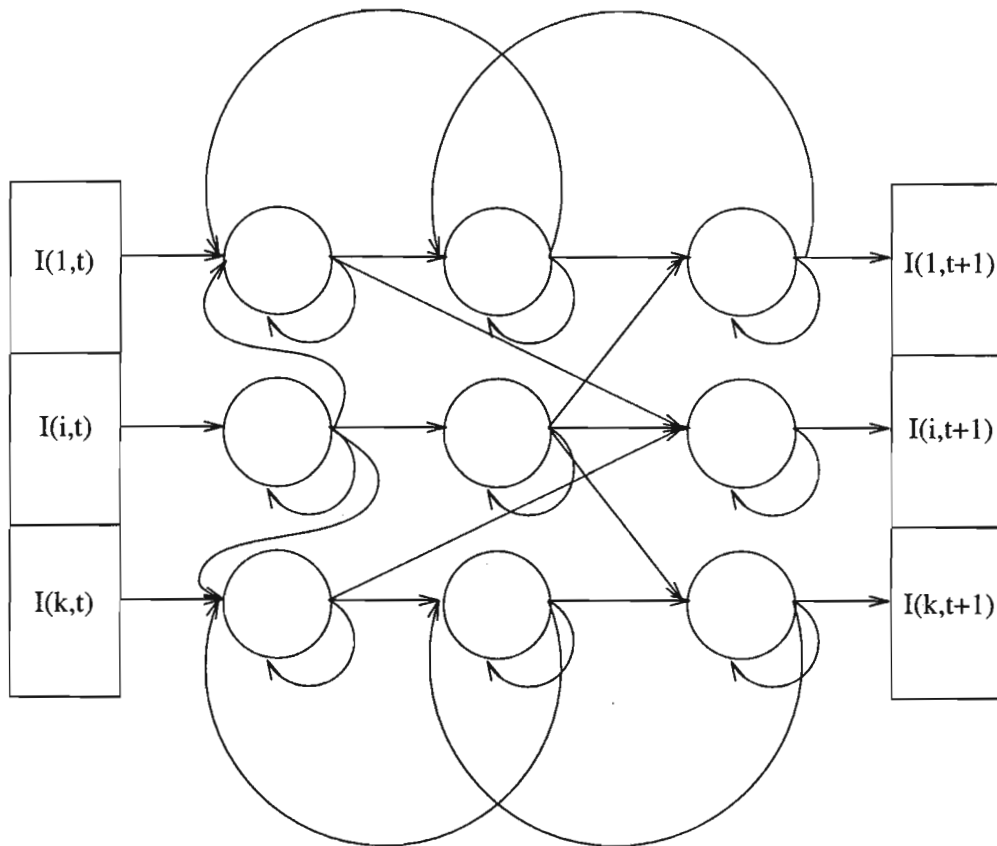


Figure 3. A network model of the shared representation. The present cochlear activation feeds into a fully interconnected network, which is trained to predict the next cochlear activation.

Human Listeners

No training	Ten Voices	=	Four Voices
			^
Training	Ten Voices	<	Four Voices

Generalization Model

Pretraining weights	Ten Voices	\geq	Four Voices
	v		^
Post-training weights	Ten Voices	<	Four Voices

Figure 4. The difference between human listeners and the generalization model's predictions. The 'ten voices' data sets represent unfamiliar speakers; training reduces performance for the model but not for human listeners.

In order to achieve an advantage for the familiar voices, the network must actually 'forget' what it learned about the initial pre-training set of voices. In particular, the pre-training regime loads the network with primarily word-dependent weights. These weights allow the network to predict what the next cochlear output will be based on the previous one under the phonetic constraints of the input. Change of the similarity space sufficient to result in enhancement of familiar voices is accomplished by loading into the network weights which reflect the transitions specific to those voices. These weights then do not support valid generalizations to words spoken by other voices. In contrast, human learners are able to maintain their existing fluency with unfamiliar speakers while acquiring knowledge about new voices.

Transfer

A transfer model supposes that familiarity with a speaker's voice increases sensitivity to the information about linguistic content. In essence, SI requires the ability to detect what is constant over the utterance, while LP requires sensitivity to what is changing rapidly. Since what is constant and what is changing in the environment are both projected in a time-varying acoustic waveform, detection of either is analogous to being able to predict a portion of the incoming signal. The residual, unpredicted portion contains both noise and information about the other object/event.

Insert Figure 5 about here

The transfer model depicted in Figure 5 supposes the following organization of speech processing. The ear filters the speech wave, resulting in a time-varying neural power spectrum map (the EM or expectation map). The activation of the map is determined by the current auditory nerve activation, the recent history of activations, as well as the listener's expectations of future activation. The current state and history of this initial, shared representation of the signal partly determines the activation of maps representing listeners' judgments about all objects of speech perception. Improvement in any secondary process results in both improved use of the information present in the shared representation, and increased fidelity in that representation.

There are two ways of viewing this model in practice. Considering the system as it functions, it forms three mutually-cascaded dynamical systems. Two dynamical systems are cascaded when the state variables of one system are parameters in the next. Mutual cascades occur when the state variables of the second system in turn serve as parameters of the first. The alternative approach is to consider the system as it learns, and here a connectionist approach is perhaps more useful. Both views of this theory are presented below.

Mutual Cascaded Dynamical Systems

Before the signal first enters the system, the expectation (or shared representation), LP and SI map each have a particular state or pattern of activation. The location of that state, which can be thought of as a particle, is determined by the attractor layout of the map. Consider first the expectation map, if it were isolated from the other two. Once speech begins to arrive, the state of the system begins to change. Acting as a perturbation, the signal is capable of moving the state away from an attractor, against the gradient of the map's potential. One proposed characterization of the shared representation is that it is a phonetic representation. If that is so, then we can imagine that the map contains attractors at allophone 'locations'. The signal then consists of a force which moves the state of the map from one such location to another.

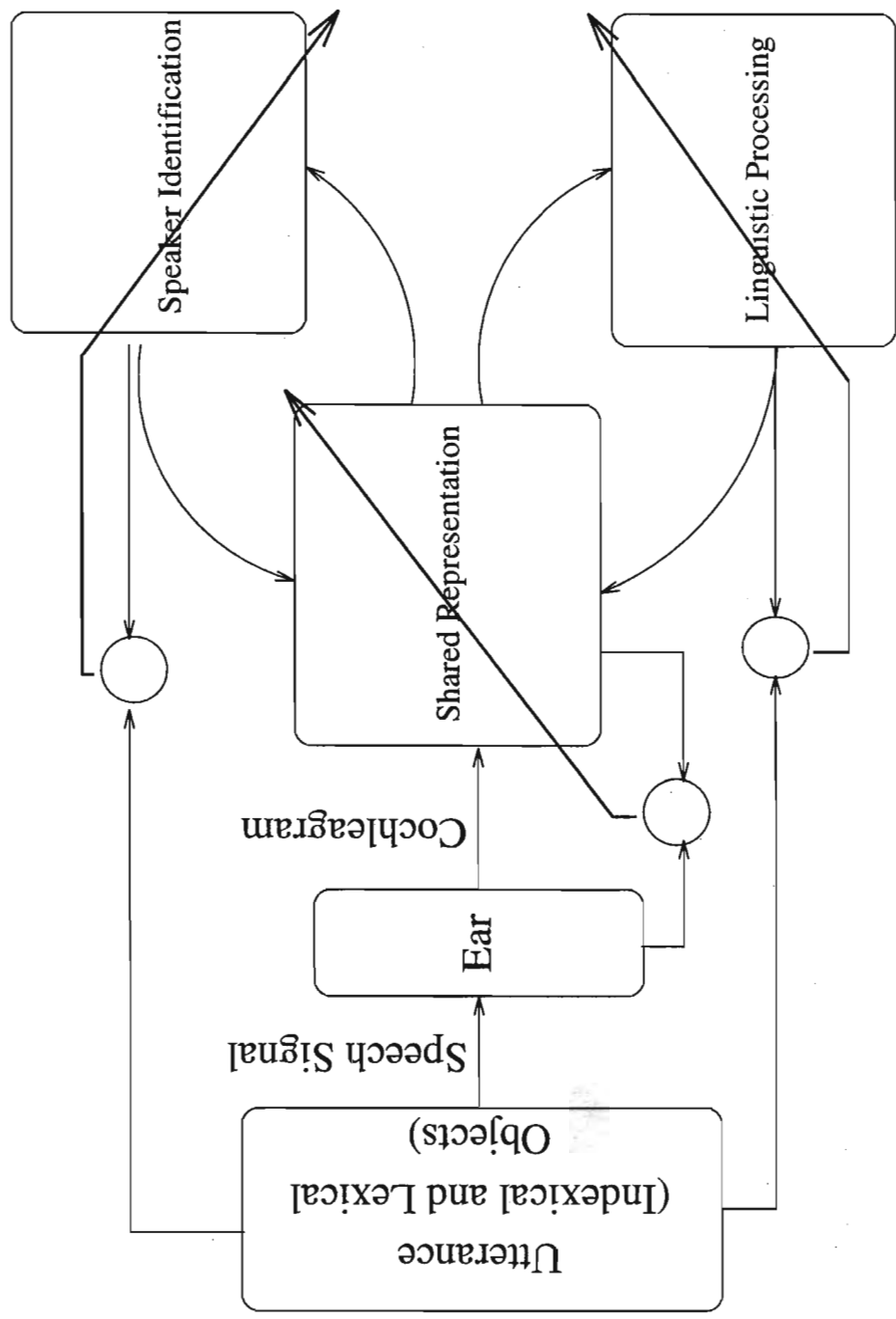


Figure 5. The transfer model. Training influences all processes, and all processes feed error back to the shared representation

Movement of the state of the expectation map from one allophone to another ought to be governed not only by the signal history, but also by the transition probabilities of allophones. To that end, the current state of the map must change the surface over which it moves. Imagine that one allophone is commonly followed by another. Once the state moves to the first location, its movement to the second ought to be easier than if it were in a less common location. This can be accomplished by using the present state of the system as a parameter, by making the layout of the map dependent on its current activation.

Allophone transition probabilities are certainly not fixed across individual speakers and between different linguistic contexts (words or phrases). This being the case, these two sources of information ought to have a strong influence on the trajectory of the activation of the expectation map. It is also precisely these sources of information which we are most interested in, theoretically. So it is natural to extend the expectation map concept by linking it to two similar maps, one which represents the particulars of the speaker, and the other the particulars of the linguistic context.

These maps are functionally similar, so consider the linguistic map first. Its activation represents the present judgment of the listener about the identity of the present word being spoken. This is clearly dependent both on the phonetic history of the signal and the constraints of the language. The phonetic history exerts its influence by reshaping the attractor layout of the LP map. Regions of the map corresponding to candidate words should be more easily activated and more easily transitioned to and from than less likely words. Because transition probabilities are determined by word-level factors as well, as in the TRACE model (McClelland & Elman, 1986), then the activation of the LP map must serve additionally to parameterize itself.

As long as activations are a function of a linear combination of inputs and weights, such as the sum of their products, then changes in input and weight have the same effect. Reparameterization of an expectation map (a group of recurrently connected units) is thus as simple as changing the activation passed between groups. The distinction between parameters and state variables is purely a distinction of perspective, rather than function in this case.

Beyond transitions between words at the LP map, word-level information influences our expectation of what phonetic segments are likely to be upcoming. Thus, the state of the LP map (its activation) must parameterize the shared representation in order to increase the likelihood that more expected allophones will be perceived. The LP map is cascaded from the shared representation, which in turn is cascaded from the SI map. The same relation holds *inter alia* for the SI map, where facts about the speaker (their vocal tract characteristics, etc.), are partially determined by and in turn help determine the history of perceived allophones.

To summarize: the trajectory of a particle across a surface serves as the input to a decision process (or is a decision process), whether for speaker identity in the SI system or linguistic properties in the LP system. The location of the EM particle is the current activation of the EM map. The change in location of the particle is under the control of processes of short term adaptation. It is also a representation of the extrinsic dynamics acting on the attractor layout of both the SI and LP maps. These maps in turn have activations, which take the form of a particle moving over their potential surfaces. The location of these particles in turn changes the dynamic of the EM layer. So familiarity is a long-term change in short-term adaptation. The coupled systems continue to change both their activations and their potentials throughout the processing of an utterance.

Multiple Recurrent Networks

From a network perspective, we can think of each model component as a recurrent network. In a recurrent network, the activation of the each node comprises the state variables, and the weights between nodes are the parameters. Connections between maps must involve using state variables at one level to change the parameterization at the next. One way to do this is indirectly. By simply using the activation of one map as input to the next, the effect of the signal input will be altered. This changes the movement of the particle along the surface not so much by changing the surface layout as by changing the inertial properties of the particle. Alternatively, the activation of one map could directly modulate the weights of another (or itself). This wholesale re-mapping of the input signal to a new pattern of activation essentially alters the topography of the surface itself.

The weights of the SI and LP networks are trained to minimize error in the detection of speaker-specific properties and linguistic structures, respectively. The outputs of the EM (expectancy map) network on the other hand serve to modify the current weights or activations of the SI and LP networks. The SI and LP networks similarly modulate the weights or activations of the EM network.

Consider only the indirect (inertial) modulation case. For exposition, imagine a case where each map has only one node, and there is only a single input from the cochlear map. The EM node has four inputs: one is from the cochleagram, one is from the node itself, one is from the SI map node and one is from the LP map node. It is the joint activations of these inputs which together determine the activation of the EM node. The weights on each input are trained both directly, and indirectly. Direct training is by way of comparing the EM node's output (its prediction of the next cochlear input) to the next input. The indirect training is by way of error propagated back from the SI and LP maps. These maps are trained to identify the identity of the speaker and linguistic content of the utterance, respectively.

The expectation map preserves information necessary for predicting the cochlear input. This information is also of use for determining speaker identity and lexical content. Perceptual learning involves increasing the fidelity of the expectation map's predictions, which might in itself increase intelligibility or identifiability. This increase in fluency should be evident in, for example, speeded response times to familiar speakers in noisy contexts. Identification training highlights those aspects of the cochleagram which are speaker specific. These specificities reduce the uncertainty of the prediction, and provide additional information for the lexical process. Under either the network or dynamical conceptualization, the model makes the same predictions. Namely, long term memory for voices and words act to shape the expectations of the listener, and expectations shape the perception of the current input sounds. Expectations can also change as a result of the immediate history of the signal.

CONCLUSIONS

We hypothesize that the ability to identify a speaker by their voice acts on interpretability in a manner analogous to short-term adaptation. That is, familiarity should result in improved interpretability in the same contexts which adaptation does, and that when combined the two should act equivalently to increased short term exposure to the voice.

The dynamics of adaptation suggest that word recognition is faster and more accurate the more that is known about the speech signal. Adaptation to a new voice takes time, so that recollection of a speaker's attributes (through speaker identification) can save time and/or increase accuracy. Through

methods such as gating, perceptual identification in noise, lexical decision and naming, the effects of foreknowledge of indexical features on lexical processing can be investigated.

In summary, the transfer model of the speaker familiarity effect provides a better explanation than does the generalization model. The transfer model also makes testable predictions about the facilitation of lexical processing given information relevant to adaptation to different acoustic features and different time scales.

References

- Halle, M. (1985). Speculations about the representations of words in memory. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 101-104). New York: Academic Press.
- Joos, M. A. (1948). Acoustic phonetics. *Language*, **24**, 1-136.
- Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. In *Proceedings IEEE Transactions on Acoustics, Speech, and Signal Processing*.
- McClelland, J. L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.
- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 309-328.
- Remez, R. E., Fellowes, J. M., Pardo, J. S., & Rubin, P. E. (1993). Voice recognition based on phonetic information. Paper presented at the 34th annual meeting of the Psychonomic Society, Washington, D.C.
- Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 915-930.
- Slaney, M. (1988). Lyon's cochlear model. *Apple Technical Report #13*, Apple Corporate Library, Cupertino, CA.
- Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalisation as demonstrated by reaction times (*Report on Research in Progress in Speech Perception No. 2*). Belfast, Northern Ireland: Queen's University of Belfast.
- Van Lancker, D. R., Cummings, J. L., Krieman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex*, **24**, 195-209.