RAW: A REAL-SPEECH MODEL FOR HUMAN WORD RECOGNITION

David van Kuijk (1+2), Peter Wittenburg (2), Ton Dijkstra (3)

(1) Department of Language and Speech, University of Nijmegen, The Netherlands
(2) Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands
(3) Nijmegen Institute for Cognition and Information, University of Nijmegen, The Netherlands
Snail-mail: P.O. Box 310, NL-6500 AH Nijmegen, The Netherlands
E-mail: kuijk@mpi.nl

ABSTRACT

In recent years computational models have become more and more important in testing processing mechanisms assumed to underlie human spoken-word recognition. Models like TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) have given us much insight in the effects of, for instance, competition between words in the mental lexicon and the use of lexical information during word recognition. However, these models neglect the effects of coarticulation and variability over time by using mock speech instead of real speech input. Here we describe a new connectionist model for spoken-word recognition which differs on a number of points from other models, in that it takes real speech as input, is based on a new architecture for the representation of time, and can adapt its own weights. Simulations with the model accurately reproduce some important effects found in human word recognition. However, the representations of words in the model and the implementation of the frequency effect should be investigated more thoroughly.

1. INTRODUCTION

Almost without problems, listeners recognize ten thousands of words every day. Human word recognition therefore seems to be a very ordinary activity. However, from another perspective the ability to recognize a word is quite amazing. Listeners are able to retrieve a particular word well within half a second from a lexicon that is estimated to contain 50.000 items or more. To do so, they must take into account the complex nature of the speech signal. Speech is variable over speakers and over time (every utterance of a word, even by one speaker, is unique), and it is difficult to segment into words because it is continuous and contains mutually dependent segments (coarticulation). These properties make it difficult to build machines that recognize speech, and they challenge our inventivity to understand how humans *can* perform this task so well.

Psycholinguists investigate the word recognition process in humans to better understand the mapping of the incoming acoustic/phonetic information onto the words stored in the mental lexicon. McQueen and Cutler (in press) provide a broad overview of many of the known effects in this area. During the last few years, computational models such as TRACE (McClelland and Elman, 1986) and Shortlist (Norris, 1994) have tried to integrate such psycholinguistic evidence concerning word recognition. However, these models did not use real speech input. Instead, they took mock speech, which is based on a phonetic transcription of words as input. In doing so, human word recognition is reduced to a relatively simple string matching.

Mock input assumes that the speech signal is first mapped onto phonemic categories before the lexicon is accessed. It ignores all nuances in the human recognition process that depend on the signal. According to a phonemic transcription, a word sequence like *ship inquiry* has the word *shipping* fully embedded in it. However, in real speech the two spoken sequences *shipping* and *ship inq...* are acoustically different in several details. Thus, a model based on real speech will to some extent behave differently from one that uses phonemes in the input sequence.

Therefore, the RAW-model (Real-speech model for Auditory Word recognition) was designed to serve as a starting point for a simulation lab which combines the use of real speech and the implementation of current psycholinguistic knowledge. The model intends to (a) adhere to the constraints defined by psycholinguists as much as possible, (b) use real speech as input, (c) store temporal patterns in a plausible way (relative to, e.g., TRACE), and (d) allow later extensions to account for the use of prosodic information and improved attentional and incremental learning mechanisms.

In the design of the model we explicitly chose *not* to use current main-stream Hidden Markov Model-based, recurrent network-based, or hybrid techniques from the world of automatic speech recognition. The rationale behind this choice was that these techniques do not provide a good basis for simulating psycholinguistic issues such as incrementally building up lexica and active competition between words. Furthermore, these architectures are not open and flexible enough to allow easy introduction of extra knowledge sources like prosody. An overview of some major limitations of these systems and suggestions for new ideas can be found in Bourlard (1995) and Wittenburg, van Kuijk, and Behnke (1995).

2. THE ARCHITECTURE OF RAW

Like earlier models, RAW relies on a hierarchical approach. Besides a preprocessing step, RAW incorporates a phonemic and a word layer. The preprocessing of the speech signal results in a number of speech vectors which form the input to a phonemic map (p-map) yielding typical activity distributions for each vector. Word neurons in the word map (w-map) sum up the activity distributions over time, leading to an activity distribution in the

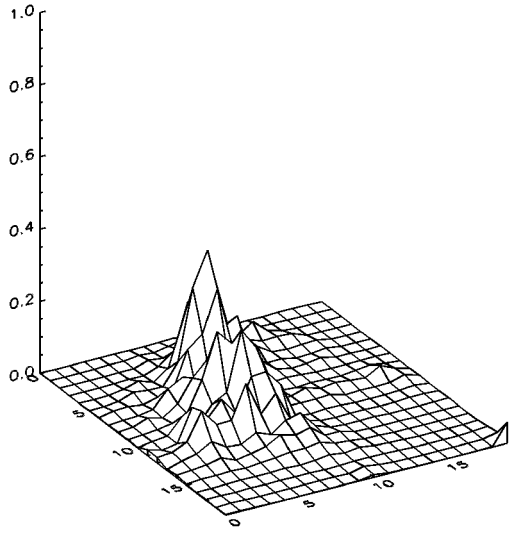

Figure 1: The activation pattern for the /i/ in *sister*. Each of the crossings in the grid represents a neuron in the p-map, and the height of the crossing represents the activation of that neuron.

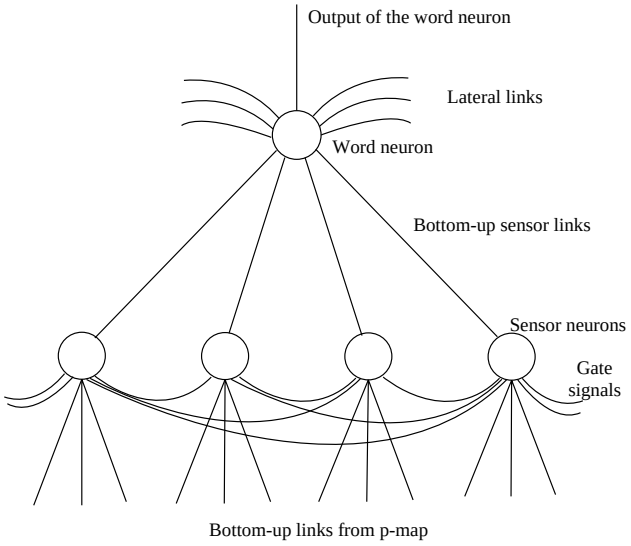

Figure 2: The chain of neurons storing the pattern of one word. It exists of a sequence of sensor neurons which are connected to the p-map via excitatory bottom-up links. Activations arise in the sensor neurons which are passed on to the word neurons by means of a special gating mechanism.

w-map as well. Competition between the word-neurons (which is suggested by psycholinguistic findings) can be simulated with the help of lateral inhibitory links, or by the implementation of a recognition rule.

2.1. Pre-lexical Processing

In automatic speech recognition mainly two techniques for preprocessing are used. The first is RASTA-mel-cepstra (Hermansky and Morgan, 1994) which yields robustness against variations in channel characteristics, and the second is Bark-scaled filter bank preprocessing (Hermansky, 1990). We used the second technique because it is simple, and because we did not have to cope with largely varying channel characteristics since our speech data were all recorded under identical circumstances. Our 16-dimensional vectors were derived from acoustical filter-bank processing based on 17.5 ms windows with a stepsize of 8.75 ms. The spectra were further preprocessed by energy normalization and noise filters.

The p-map is necessary for context-dependent decomposition of the highly modulated segmental information in the speech signal. The ultimate goal of the p-map is to generate, for every incoming speech vector, an activation distribution characterizing the speech segment represented by that vector in its acoustic/phonetic context. To do so we used a self-organizing feature map (Kohonen, 1988), which carries out a data and dimensionality reduction of the input space, while at the same time preserving the relations of similarity between the input vectors. As an example; in the trained map the activation patterns for different realizations of a /th/ will be almost identical in shape and position on the map. The activation peak for /s/, which is acoustically very similar to /th/ will arise near the peak for /th/, but it will have a different shape. For phonemes which are acoustically more different from /th/ the activation bubble will be clearly separable from the /th/-bubbles. Figure 1 shows the activation pattern for the /i/ in the word *sister*. The implementation of the use of temporal context in the p-map is still being studied.

2.2. The Word-Map

The word-map (w-map) uses the activity distributions in the p-map over time to store the acoustical properties of each word. During recognition, the word neurons must accumulate matching information, but also block incoming activation when the match between stored word and input is bad for a part of the left-hand context. So the input *think* must to a certain extent activate both the lexical entries *think* and *thinking*, but the input *king* must not activate *thinking* because the left-hand context of that entry (*thin*) was not present in the input. The word neurons must also possess the capacity to store the inherent timing of the spoken words. In RAW temporal information about a word is stored in the word neuron in a sequence of sensor neurons which are connected via so-called gate signals (see Figure 2).

Every sensor neuron is sensitive to a certain pattern in the p-map. The **potential** of the sensory neurons is computed as a quadratic form difference between the vector described by the afferent weights and that described by the relevant activity distribution of the p-map neurons they are linked to. The **activation** of each sensory neuron is calculated by multiplying its potential with the effective gate value of that neuron at that moment. The potential of the sensors is context-independent (so during the input *king* the sensors in the second syllable of the lexical entry *thinking* will also

have high potentials), but the activation of the sensors is context-dependent because the gate signal is context-dependent.

The gate signal is the most important construct for representing the sequential structure of a word. During recognition it is computed at each time slice as a function of the match between the word represented by each word-neuron and the input so far. If the gate of a sensor is not open, that sensor can not be activated. The gate of the first sensor of each word is always open, so that a word can always be activated from its beginning. If the mismatch between input and word neuron is too high all gates of that word close. As a result no more bottom-up activation reaches the word neuron, and so it is thrown out of the competition for recognition. The gate function is flexible enough to allow RAW to catch up with speaking rate variations, omissions and other small distortions of the speech signal. In fact, the gate signals have to fulfil a job similar to that of dynamic programming algorithms in HMM-based systems.

The activation of the word neuron is a function of all sensor activations at any moment in time and the accumulated activation from earlier time steps. Figure 3 shows the activations over time in the word-map for the input *sister*.

The bottom-up links between the p-map and the sensory neurons are trained in two stages: a bootstrapping phase and a fine tuning phase. In the bootstrapping phase, one good articulation is used to initialize the sequence which results in an excellent match of this specific token. During the supervised fine tuning phase the links are adapted such that they focus on the salient and discriminative aspects of the different variants produced.

The bottom-up links between the sensory neurons and the word neurons are constant within each word, but differ between words. For high-frequent words these links are stronger, so that a high-frequent version of a lexical entry reaches a higher word activation than a low-frequent version on the basis of the same acoustical input.

A word is considered to be recognized when its word activation is sufficiently higher than that of all other words in the lexicon. In this way competition between words is modeled in a simple and efficient manner.

The complex dynamics of the model are described in more detail in Wittenburg, van Kuijk, and Dijkstra (1996).

3. SIMULATION RESULTS

For the simulation results we constructed a lexicon with 32 word entries, introducing particular relationships between the words. A number of entries were chosen that were very similar in phonological form, e.g., words like *thanking*, *thinking*, and *sinking.* Furthermore, different types of embeddings of words in other words were introduced, such as *think* and *king* in *thinking,* and *tree* in *treaties*. Each word was spoken four times by a female speaker. Since the fine-tuning is not implemented yet, the model was trained with one token of each word. Of the 128 tokens 119 were recognized correctly, which is sufficient for the psycholinguistic simulations. We analyzed the behavior of the model with respect to a number of psycholinguistically relevant factors: Uniqueness Point, Cohort Size, and Word Frequency.

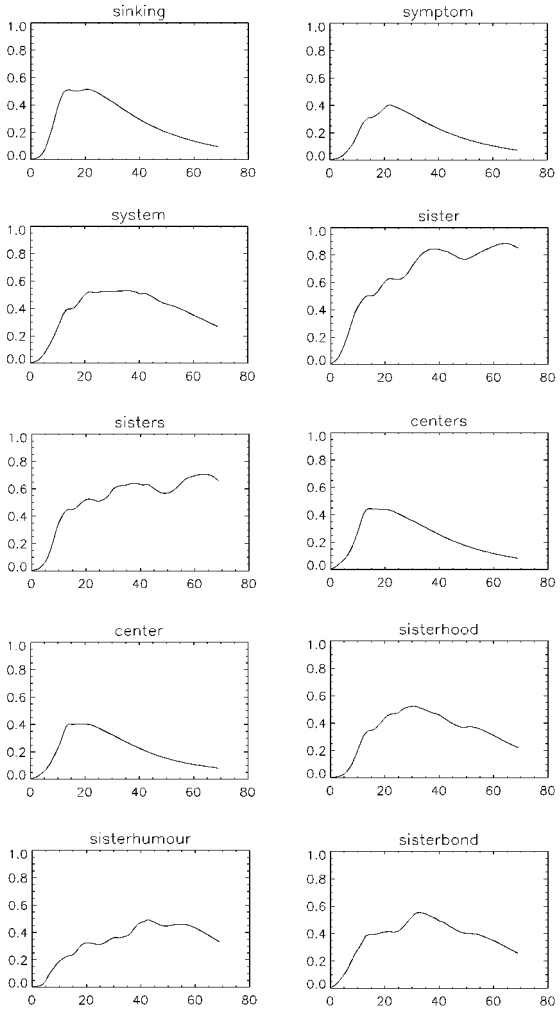

Figure 3: Activations over time of the words in the lexicon as a result of the input *sister* The x-axis shows the time-dimension in terms of time slices, the y-axis the activation.

The uniqueness point (UP) of a word is that point in the signal at which the word becomes unique with respect to all other entries in the lexicon. This point has been shown to be a valuable predictor of reaction time in human listeners. In our lexicon, an almost equal number of words had an early (68) and a late (60) UP. Measured from the beginning, the average recognition time of words with early UPs was about 51 time slices, and that of other words about 70 time slices. This indicates that words with early UP are recognized earlier, a finding which seems comparable to that for humans.

The number of words that is still consistent with the speech signal at a particular moment in time is called a *cohort*. The effect of small (< 4) vs. large (> 3) word-initial cohort sizes was analyzed. Words were grouped in two categories, dependent on the number

of word candidates after the first phoneme. The 72 word tokens with a small cohort size were recognized after on average 54 time slices, while the other group was recognized after on average 69 time slices. Thus, words in a small cohort were recognized earlier than those in a large cohort, similar to human word recognition.

Many psycholinguistic experiments show that human listeners recognize high frequent words faster than low frequent words. The RAW model exhibits a comparable effect when frequency is simulated by allowing the activations of high-frequent words to increase faster than those of low-frequent words, but the effect is very small. Furthermore, this implementation of word frequency also led to some side effects unknown for human subjects. For example, if the word *flashing* is assigned a higher frequency than *flushing*, the word recognized is almost always *flashing*. Clearly, it is not desirable that word frequency overrides bottom-up acoustic information in this way.

4. DISCUSSION AND CONCLUSIONS

The RAW model shows a potential to simulate a number of psycholinguistic effects on word recognition and therefore to serve as a simulation and theorizing tool. The dynamics of the network provides a promising basis for further investigation. It performs better than TRACE in that it may pick up early deviations in the pronunciation, for instance, the difference in the way the first vowel is spoken in *tree* and *treaties*. While this seems to be a desirable feature, more information should be collected about how and when human listeners make use of such more subtle differences in the speech signal. The current implementation of word frequency in RAW resulted in some undesirable effects. Relating word frequency and bottom-up information to the same activation function may be a main reason for this. Other plausible architectural solutions which fit better with current psycholinguistic insights have to be implemented and tested.

The speech recognition component in RAW can be improved in two ways. First, as already mentioned we are studying a better pre-lexical processing which takes the acoustic context into account. This new p-map should deliver a richer and more tunable output representation than for instance recurrent neural networks do. Second, the fine tuning phase of the training has to yield a more abstract representation for each word, which is aimed at maximizing the differences in representations between words, while minimizing those differences within a representation of a word. This discrimination can be achieved by including attentional mechanisms. In case of similar sound patterns of two different words (e.g., *flushing* and *flashing*) the w-map has to be trained such that the stored patterns yield a maximal difference. In case of strong differences between tokens of the same word, a complex graph will evolve compared to the simple sequences we have now.

After this fine-tuning is implemented we can examine whether the present implementation of the frequency effect is realistic or not. The psycholinguistic component of the model can further be improved by the inclusion of prosodic information in the model. Cutler and Norris (1988) suggest that English listeners use the weak/strong syllable distinction to guide lexical access. It has to be investigated in how far the gate signals can be optimized by using such information.

5. REFERENCES

Bourlard, H. (1995). "Towards Increasing Speech Recognition Error Rates". *Proceedings Eurospeech 95*. Madrid.

Cutler, A., and Norris, D. (1988). "The role of strong syllables in segmentation for lexical access". *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 113-121.

Hermansky, H., and Morgan, N. (1994). "RASTA Processing of speech". *IEEE Trans. Speech Audio*, 2(4).

Hermansky, H. (1990). "Perceptional Linear Predictive Analysis of Speech". *JASA*. 87 (4). 1738-1752

Kohonen, T. (1988). "The neural phonetic typewriter". *Computer, 21(3)*, 11-22.

McClelland, J., and Elman, J.L. (1986). "The TRACE model of speech perception". *Cognitive Psychology*, 18, 1-86

McQueen, J.M., and Cutler, A. (in press). "Cognitive Processes in Speech Perception". In W.J. Hardcastle and J. Laver (Eds.), *A Handbook of Phonetic Science*. Oxford: Blackwell.

Morgan, N., and Bourlard, H. (1995). *Connectionist speech recognition*. Dordrecht: Kluwer.

Norris, D. G. (1994). "Shortlist: A connectionist model of continuous speech recognition". *Cognition*, 52(3), 1212-1232.

Wittenburg, P., van Kuijk, D., and Behnke, K. (1995). "Automatic and Human Speech Recognition Systems: a Comparison". *Proceedings 3: SNN Symposium*. Nijmegen.

Wittenburg, P., van Kuijk, D., and Dijkstra, T. (1996). "Modeling human word recognition with sequences of artificial neurons". *Proceedings ICANN96*. Bochum.