Modeling Human Word Recognition with Sequences of Artificial Neurons

P. Wittenburg, D. van Kuijk, T. Dijkstra*
Peter.Wittenburg@mpi.nl
Max-Planck-Institute for Psycholinguistics, Nijmegen
*NICI, University of Nijmegen, Nijmegen
The Netherlands

Abstract. A new psycholinguistically motivated and neural network based model of human word recognition is presented. In contrast to earlier models it uses real speech as input. At the word layer acoustical and temporal information is stored by sequences of connected sensory neurons which pass on sensor potentials to a word neuron. In experiments with a small lexicon which includes groups of very similar word forms, the model meets high standards with respect to word recognition and simulates a number of well-known psycholinguistical effects.

1 Introduction

Human listeners can process speech seemingly effortless. Even degraded speech can be processed at a rate of about 2 to 3 words per second. Listeners can perform this amazing feat by relying not only on the signal (bottom-up information), but also on syntactic and semantic information extracted from the left-hand context (top-down information). However, even words spoken in isolation are recognized extremely well by human listeners, although performance is not always flawless.

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 [1] provide a broad overview of many of the known effects in this area. During the last few years, processing models have tried to integrate such psycholinguistic evidence concerning word recognition [2,3]. However, up till now these models have not used real speech input. Instead, they have used mock speech input, based on a phonetic transcription of words, to investigate relevant psychological issues (such as competition effects among word candidates). In doing so, human word recognition is reduced to a relatively simple string matching.

Mock input presumes a solution of how the speech signal is mapped onto abstract phonemic categories, and ignores all nuances in the human recognition process that depend on the signal in relation to the recognition task at hand. 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/phonetically 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 [3]), 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 explicitely chose not to use current main-stream HMM-based or hybrid [see e.g. 4] techniques from the world of Automatic Speech Recognition (ASR). The rationale behind this choice was that these techniques do not provide a good basis for simulating psycholinguistic issues such as gradual lexicon expansion 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 [5] and [6].

2 The Architecture of RAW

Like earlier models and systems, 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. The p-map can be seen as a kind of spatio-temporal filter. Word neurons in the word map (w-map) sum up the activity distributions over time, leading to an activity distribution in the 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.

2.1 Pre-lexical processing

In ASR mainly two techniques for preprocessing are used. The first is RASTA-mel-cepstra [7] which especially yields robustness against variations in channel characteristics, and the second is Bark-scaled filter bank preprocessing as suggested by Hermansky [8]. We used the second technique because it is simple, and because we did not have to cope with largely varying channel characteristics. The speech signal was multiplied by a 17.5 ms Hamming window with a stepsize of 8.75 ms. Next, the spectral representations obtained with a 512 point FFT were nonlinearly transformed to the Bark scale and finally multiplied with equidistant acoustical bandfilters [8]. These 16-dimensional spectral representations were further preprocessed by energy normalisation 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. The scalar output of supervised trained MLPs [4], TDNNs [9], or RNNs [10] as used in hybrid ASR systems is too restricted for this goal. Instead we used a self-organizing feature map [11], which carries out a data and dimension reduction of the input space, at the same time preserving similarity relations between the input vectors. For example, in the trained map the activation peaks for different realisations of a /th/ will be almost identical in shape and position on the map. The activation peaks for /s/, which is acoustically very similar to /th/, will arise near the peaks for /th/, but it will have a different shape. For phonemes which are acoustically more different from /th/ the activation peaks will be clearly separable from the /th/-peaks. The implementation of the use of context in the p-map is currently under study. Since this paper focusses on the construction of the word map, the p-map and its characteristics will not further be discussed.

2.2 The word map

The word-map (w-map) uses the activity distributions in the p-map over time to store the sequential and the spatial representation of each word. The word neuron assemblies must accumulate matching information, but also block incoming activation when no input is expected. These neuron assemblies must further possess the capacity to store the inherent timing of spoken words and to store temporal order. This is done in RAW by defining for each word at least one sequence of sensor neurons connected via so-called gate signals (see figure 1).

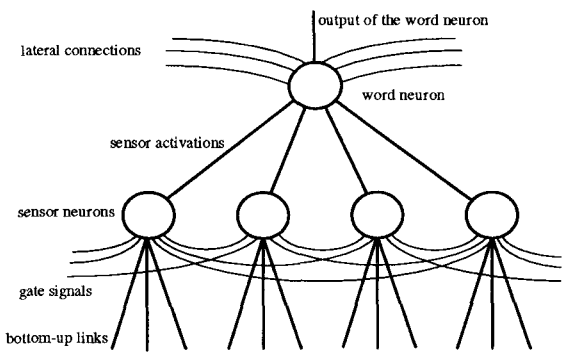

Figure 1 shows the assembly of neurons storing the pattern of one word. It exists of a sequence of sensor neurons which are connected to the p-map via excitatory connections. Bottom-up activation leads to sensor activations which are passed on by means of a special gating mechanism to the word neurons. Word neurons can compete with other word neurons via lateral links.

Every sensor neuron is sensitive to the activation pattern of only a certain spatial location in the p-map. The potential of the sensor neurons is computed as a quadratic form difference between the vector described by the afferent weights and that

$$q_{i,k}^{pot}(t) = \frac{1}{N}\sum_{n=0}^{N} e^{a1*(w_{i,k,n}-x_n(t))^2} \quad (1) \qquad q_{i,k}^{act}(t) = g_{i,k}^{act}(t) * q_{i,k}^{pot}(t) \quad (2)$$

described by the relevant activity distribution of the p-map neurons they are linked to (1), where a1 is a parameter <0, n is a neuron in the p-map, $w_{i,k,n}$ the link from a p-map location to sensory neuron k of word neuron I, x_n the activation of p-map neuron n, and N is the number of afferent links of sensor neuron k. The potential is transformed into the sensor activation by multiplying it with the corresponding gate signal (2). In doing so, a sensor activation will be high when the stored typical pattern appears in the p-map, given that the gate signal is high, i.e. if the pattern occurs at the expected moment relative to the preceding stored patterns.

The gate signal is the most important construct for representing the sequential structure of a word. At any moment during processing it represents the expectation value for the activation of the corresponding sensor neuron. Thus, it prevents the transfer of a high sensor potential, caused by a high local match, at a wrong moment in time. At the same time, the gate signal should be high when the previous input matches perfectly well with the stored pattern up to that moment. In this case the sensor potential should contribute to the global word activation. Equation 3 describes the complex dynamics of the gate potential:

$$g_{i,k}^{pot}(t+1) = g_{i,k}^{pot}(t) * (1 - dec1) + a2 * q_{i,k}^{act}(t) + a3 * \sum_{r=0}^{R} (\gamma_r * g_{i,k-r}^{act}(t)) \quad (3)$$

with a2, a3, and dec1 parameters, γ a Gaussian function, and R the scope of the Gaussian. The recursive influence of the gate values g^{pot} effectuates a slow decrease of the gate signals. The second additive term describes the influence of the local match at the previous moment, and the third term represents the sum of earlier gate values weighted by a Gaussian function. This, and the dependency on the amount of match at the preceding sensor neurons, allows RAW to catch up with speaking rate variations, omissions, and other distortions of the speech signal. The gate signals are normalized variants of the gate potentials (4).

$$g_{i,k}^{act}(t) = \frac{2}{1 + e^{-g_{i,k}^{pot}(t)}} - 1 \quad (4)$$

In fact, the gate signals have to fulfil a job which is similar to that carried out by dynamic programming algorithms in HMM-based systems. However, gate signals do not need backtracking and operate directly. The global potential of the word neuron is a summation of all sensor activations at any moment in time and the accumulated potential from earlier time steps (5) where dec2 and a4 are parameters, p_i the accumulated potential of word neuron i, and $q_{i,k}^{act}$ the contributions of the sensor neuron k. The final word activation is achieved by normalization (6).

$$p_i(t) = p_i(t-1) * (1 - dec2) + a4 * \sum_{k=0}^{K} q_{i,k}^{act}(t) \; ; \quad y_i(t) = \frac{2}{1 + e^{-p_i(t)}} - 1 \quad (5,6)$$

The afferent weights between the p-map and the sensory neurons are trained in two phases: a bootstrapping phase and a fine tuning phase. In the bootstrapping phase one well-articulated token is used to initialize the sequence of sensor neurons which

results in an excellent match of this specific token. During the supervised fine tuning phase the weights are adapted such that they focus on the salient and discriminative aspects of the different variants produced. The strengths of the weights are correlated with the contribution of the singular connections to the sensor potential.

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 /sister/ in /sisters/, and /tree/ in /treaties/. We analysed the behavior of the model with respect to a number of psycholinguistically relevant factors: Uniqueness Point (UP), Cohort Size, and Word Frequency. The uniqueness point is that point in the acoustic signal at which the word becomes unique with respect to all other entries in the lexicon. The cohort of a word is that number of words that is still consistent with the speech signal at a particular moment in time. At the UP the cohort size is reduced to one. All three factors affect the recognition process of humans.

The simulations showed a clear relation between the UP of the words in the lexicon and their point of recognition. This is consistent with findings in experiments with humans. With respect to the influence of the size of the cohort we also obtained consistent results. Words in a small cohort were recognized earlier than those in a large cohort, similar to human word recognition. The way the frequency effect was implemented in RAW, multiplying the contributions to the global word potential with a factor correlated with the word frequency, led to a clear frequency effect in recognition. Unfortunately, this implementation of word frequency also led to some side effects unknown for human subjects.

For a more detailed discussion of the simulation results and psycholinguistic findings we refer to another forthcoming paper [12].

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 [3] in that it may pick up early deviations in the pronunciation, for instance, the difference in the way the /I/ 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: (1) As already mentioned we are still looking for a better pre-lexical processing. The method has to take the acoustic context into account, like for example RNN do. On the other hand, the method should deliver a richer and more tunable output representation than that of RNN. A temporal solution will be the inclusion of the spectral and energy differences trained in separate maps. A further improvement of pre-lexical processing is expected by reducing the number of stored patterns with the help of a trace segmentation algorithm. This will give more weight to the dynamic parts of the signal. (2) The fine tuning phase of the training has to include attentional, i.e. discriminative mechanisms. In case of similar sound patterns of two words, the w-map has to be trained such that the stored patterns yield a maximal difference in activation. In case of strong differences between tokens of the same word, a new assembly of neurons will be used to store this variant. Fine tuning comes out on the one hand as modifying the representations of an existing neuron assembly or on the other hand as storing largely deviating patterns in separate neuron assemblies.

Of great importance is the inclusion of prosodic and syllabic information in psycholinguistic models. Syllable boundaries are possible moments of articulatory synchronisation. It has to be investigated in how far the gate signals can be optimized by using this information.

References:

[1] McQueen, J.M. & Cutler,A. (in press). Cognitive Processes in Speech Perception. In W.J. Hardcastle & J. Laver (Eds.), A Handbook of Phonetic Science. Oxford: Blackwell

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

[3] McClelland, J. & Elman,J.L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86

[4] Morgan, N. & Bourlard, H. (1995). Connectionist Speech Recognition. Dordrecht:Kluwer.

[5] Bourlard, H. (1995). Towards Increasing Speech Recognition Error Rates. Proceedings Eurospeech95, Madrid

[6] Wittenburg, P., van Kuijk, D. & Behnke, K. (1995) Automatic and Human Speech Recognition Systems: a Comparison. Proceedings 3. SNN Symposium, Nijmegen, NL.

[7] Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. IEEE Trans. Speech Audio, 2(4)

[8] Hermansky, H. (1990). Perceptional Linear Predictive Analysis of Speech. J. Acoust. Soc. Am. 87 (4). 1738-1752

[9] Waibel, A. et al. (1987). Phoneme Recognition Using Time-Delay Neural Networks. Technical Report TR-1-0006, ART Interpreting Telephony Research Laboratories.

[10] Wittenburg, P. & Couwenberg, R. (1991). Recurrent Neural Networks as Phoneme Spotters. In T. Kohonen et al. (Eds.), Artificial Neural Networks. Amsterdam: North Holland.

[11] Kohonen, T. (1989). Self-Organization and Associative Memory. Berlin: Springer Verlag.

[12] van Kuijk, D.,Wittenburg, P. & Dijkstra, T. (1996). A connectionist model for the simulation of human spoken-word recognition. Proceedings of the 6th Workshop Computers in Psychology 1996, Amsterdam.