USING NEURAL NETWORKS TO INVESTIGATE THE RELATIONSHIP BETWEEN SPEECH PRODUCTION AND PERCEPTION

Jörgen Pind

Department of Psychology, University of Iceland, Reykjavík, Iceland

ABSTRACT

The relationship between speech perception and speech production is a complex one. Different views have been offered on this relationship in the past. In particular, the question of the extent to which perception mirrors production has been contested. For this relationship to be fruitfully investigated it would be advantageous to adopt some objective means of deriving perceptual hypotheses from production data. This paper shows that simple neural networks can be used for this purpose. Two simulation experiments will be presented and compared with experiments dealing with the rate-dependent perception of Voice Onset Time (VOT) and of duration as a cue for vowel and consonant quantity. Results show that a neural network containing no hidden units is able to predict the result of a perception experiment dealing with vowel and consonant durations as cues to quantity, even at variable speaking rates, but fails to predict the results of an experiment dealing with the rate-dependent perception of VOT.

1. RELATING PRODUCTION AND PERCEPTION

It would appear to stand to reason that listeners perceive speech from the acoustic signal and therefore that one should expect a tight fit between the acoustics of speech and speech perception. As is well known, however, theories of speech perception have often questioned this fit, and been struck by the fact that speech acoustics and perception sometimes seem to “go their separate ways”. Much was made of this in early work at the Haskins Laboratories, leading Liberman [5] to pose the question “When articulation and soundwave go their separate ways, which way does the perception go?” His answer was that “perception always goes with articulation.” This sentiment was later reflected in the motor theory of speech perception [6]. This view, though, has by no means gone undisputed. Thus Fant [3] made a forceful claim to the effect that “articulation and sound waves never go separate ways”.

One factor which has made it difficult to resolve the question as to the tightness of the fit between acoustics and perception, is the fact that speech is highly variable. This variability is caused by numerous factors having to do with differences in voice quality, the effect of coarticulation, different utterance rates and so on.

Speaking rate is a particularly important cause of variability with respect to temporal speech cues. A well-known cue of this kind is Voice Onset Time (VOT) which is a measure of the time between oral release of a stop consonant and the start of vowel voicing. In languages such as English and Icelandic — which have two series of stop consonants, /bdg/ and /ptk/ — the difference between these two categories is conveniently captured by the VOT metric. The stops of the first series have short-lag VOTs of perhaps 10–25 ms, the latter series show much longer values of VOT, typically longer than 60 ms.

It has long been realized that a temporal speech cue such as VOT is susceptible to extraneous influences, caused e.g. by changes in speaking rate. These changes have nothing to do with the linguistic message which is conveyed by VOT and the question arises as to how the listener will cope with the problem of keeping the linguistic and the non-linguistic factors apart.

Production studies of rate effects have commonly shown that with slower speech the contrast between temporally contrasting phonemes increases. Miller and Baer [8] showed this for transition durations as a cue for the [b-w] distinction, and Miller, Green and Reeves [9] reported a similar finding for VOT. In both cases the greatest change was seen in the segment cued by the longer transitions [w] or the longer VOTs (aspirated stops). Thus in the case of VOT, slower utterance rates will lengthen the VOT for aspirated stops more than the VOT for unaspirated stops, increasing the separation of the two categories. This finding also clearly emerged from a study of the effect of speaking rate on VOT in Icelandic [12].

Research on the perception of VOT has shown that listeners appear to be influenced by the perceived speaking rate. In the case of VOT phoneme boundaries separating short and long lag categories of stops move to longer values of VOT when the overall duration of the syllables is lengthened [10, 16]. Thus a study of American English speakers found that the phoneme boundary in a /bi-pi/ continuum was located at 43.9 ms when the syllables were 325 ms long, but at 35.6 ms when the syllables were 125 ms long [10]. Such shifts in the location of the phoneme boundaries as a function of the perceived rate are commonly thought to reflect a “taking-into-account” type of perceptual mechanism — in this case appropriately termed “rate-dependent perception”.

This notion of “taking-into-account” is well-established in perceptual psychology in general, where it has been used e.g. to account for the perceptual constancies [2]. In the case of visual perception it is thus commonly found that the perception of size depends on the perceived distance of the object from the viewer. Size constancy is, at least over not too large distances, remarkably robust.

Assuming that a similar type of mechanism is operating in phonetic perception the question arises as to the extent of the taking-into-account. Will the listener compensate perfectly for
changes in speaking rate? This question has received much less attention than it deserves though some attempts have been made to explore quantitatively the degree of perceptual normalization in rate-dependent perception.

In their study of the behavior of VOTs under transformations of rate, Miller et al. [9] introduced the notion of a “variable optimal boundary” Miller et al. calculated this variable optimal boundary by sorting their VOTs measured at various utterance rates into 50 ms bins and calculating hypothetical phoneme boundaries that would optimally separate the two categories of unaspirated and aspirated stops for each bin. Miller et al. showed that this optimal boundary increased with slower utterance rates (as judged from the overall duration of the syllables) from 14 ms of VOT in 100–149 ms long syllables to 52.5 ms of VOT in 650–700 ms long syllables. The perceptual shifts are typically considerably more limited, as shown by the findings of Miller and Volaitis [10].

It is clear from the studies referenced so far that the question of how perception mirrors production can be fruitfully investigated by attempting to predict the results of perceptual experiments assuming that the production data determine the results of perceptual experiments. Various methods can be used to derive such calculations. The purpose of this paper is to argue for the usefulness of using simple neural networks for this purpose. This will be illustrated with reference to two experiments where stimuli for perception experiments were carefully modeled on production data.

2. NEURAL NETWORKS

Neural networks have revolutionized the study of pattern recognition in the last decade or so. They have also had great influence on modeling work in psychology and, to a lesser extent, linguistics. It is the position of this paper that neural networks can also be a useful tool for probing the relationship between speech production and perception.

Neural networks come in various sizes and shapes and use a multitude of different learning algorithms. One particular kind of neural network is that popularized by Rumelhart, McClelland and co-workers [15] which makes use of the “backpropagation” algorithm for training the network. A backpropagation network contains a number of connected nodes in a layered net. A network would typically have a layer of input nodes and a layer of output nodes. It might also contain a “hidden” layer of nodes interposed between the input and output layers. The task for the network is to learn some mapping between input and output. The input might be a set of English verbs in the present tense, the output the corresponding set of past tenses of the English verbs. The network has only one way of learning the mapping of input to output and that is by modifying the weights connecting the nodes of the network. These weights are automatically set by the learning algorithm used by the network, such as the backpropagation algorithm. Backpropagation is a supervised learning algorithm which means that the network will learn by comparing the output of the net to the desired output. If these do not agree an error is calculated from the difference between the actual net output and the desired output. This error is then “backpropagated” through the net leading to adjustments of weights between nodes. The backpropagation net learns by “gradient descent”, i.e. by attempting to minimize the error by moving downhill in an “error landscape”. Backpropagation is the most popular learning algorithm employed in current neural networks.

Numerous simulators are available which make it easy for researchers to design their own neural networks. One such simulator is the tlearn simulator developed by Jeffrey Elman, Kim Plunkett and associates and recently described in detail in a number of accessible publications [1, 7, 14]. This simulator, available for both Macintosh and Windows, is extremely easy to use and has been used in the following simulations.

3. FEEDING SPEECH DATA TO A SIMPLE NEURAL NETWORK

The tlearn simulator will be used here to answer the following question: Based on production data, say on the duration of vowels and consonants in stressed syllables in Icelandic (these being either of the type V:C or VC:), what predictions can be made from the production data about the optimal way of distinguishing the two categories of speech segments?

A simple neural network, shown in Figure 1, will be used to make this prediction. This will be done by feeding measurement data to the network (the input signal) along with a description of the intended category, V:C or VC:. The intended category forms the desired output signal, this being either set as 0 or 1. The network is then trained on this input-output relationship.

![Figure 1. The structure of the neural network. Measurements of speech segment durations are fed to the two input nodes while the output node is trained to make the correct binary classification of the input measurements.

Training the network in this manner leads to the weights of the network being set. After successful training the network will have partitioned the possible input space, i.e. of vowel and consonant durations, into two categories corresponding to the two syllables, V:C and VC:. This partitioning can be explored by feeding vowel and consonant durations ranging over all the stimulus space and measuring the activation of the network at each point. This is illustrated in Figure 2. This figure shows, in the zero plane using white and black circles, actual measurement values for vowel and consonant durations in 720 syllables. The
Figure 1. The mesh plot shows the network activation for durational values of consonants and vowels ranging from 50 to 300 ms. Superimposed on the training data is a line that shows the phoneme boundaries obtained by the network after training. This, as expected, nicely separates the two clusters of the training data, almost without error. The production data appear on the whole to be well separated in the scatter plots, to such an extent that it is possible to calculate a straight line which will separate the two classes of segments in both cases with quite acceptable accuracy. We thus want the network to calculate this optimal separating line. Having no hidden units, the network will be forced to treat the classification problem as linearly separable and will find the best line which distinguishes the two categories [4].

4. TWO SIMULATIONS

The simulations presented here will be based on a recent paper which compared the rate-dependent perception of VOT and quantity in Icelandic [12]. As previously mentioned, some rate-dependent effects are usually found in the perception of VOT though these effects are usually weaker than would be expected on the basis of production data and assuming a perfect “taking-into-account” type of perceptual normalization. Earlier studies dealing with the perception of quantity in Icelandic [11] had indicated that the extent of perceptual normalization seen with changes in speech segment duration is very close to that which would be expected on the basis of the production data. This result had been used to argue for the existence of a higher-order invariant for quantity through a V/rhyme ratio [11].

In the production study, four participants were asked to read the words gala [ka:la] “to yell”, galla [ka:la] “overalls, acc. sg.”, kala, [k’a:la] “suffer frostbite” and Kalla [k’a:la], familiar form of the name “Karl”, accusative. These words are distinguished by the VOT of the word-initial stop, this being either /g/ or /k/, and by the quantity of the initial syllable, one of these having the structure V:C, the other the structure VC. The participants read the words a number of times at five different utterance rates, from very slow to very fast. A total of 400 tokens were measured. Figure 3 shows the production results (in the background, using circles) for VOT and Figure 4 the results for vowel and consonant duration (here a subset of the data is presented, i.e. all words starting with /g/). The former figure shows that VOT changes with increases in syllable duration, as speech rate is slowed VOT for the long-lag stops, i.e. /k/, gets longer. VOT for the short-lag stops changes very little as a function of speaking rate. The results for vowel and consonant durations show an interesting pattern. Both vowel and consonants show duration changes as a function of speaking rate, but clearly it is the phonologically long segment that in each case shows the greatest changes in duration.

The network was run in two separate simulations. In the VOT simulation input node 1 was fed with the duration of VOT, input node 2 with the syllable duration (400 measurements in all, data shown in Figure 3). In the quantity simulation the network was fed on measurements from 200 syllables, i.e. all those starting with /g/ (data shown in Figure 4). Initially the nets were trained for 200 epochs, using a learning rate of 0.7 and a momentum of 0.8. The network quickly converged on a solution to the quantity problem. However, additional training was necessary before the VOT network had stabilized (here results after 10,000 epochs are reported).

After the networks had been trained on the production, the networks were used to classify stimulus continua corresponding to the continua used in the perception experiments of Pind [12] — experiments 2 and 3. This was done by feeding the network with measurement values corresponding to those used in the stimulus continua used in the perception experiments. In the case of VOT, though, the stimulus continua had to be extended, as described below.

There were three VOT continua, /gala-kala/, corresponding to a fast, normal and slow utterance rate (with initial syllable durations of 280, 380, and 525 ms respectively). In these continua VOT was varied from 0–100 ms. (In the perception experiment the VOTs varied from 0–60 ms, but this was not a sufficient range for the network). Results are shown in Figure 3. For comparison the results from the perception experiment are also shown. It is clear that the slope of the line connection the phoneme boundaries found by the neural network on the one hand, and the phoneme boundaries of the human listeners, are widely divergent. The neural network greatly overestimates the amount of rate-dependency actually observed in the perception experiment, which was actually quite modest. Changing the syllable duration from 280 to 525 ms moved the phoneme boundaries for listeners from 36.4 to 37.8 ms. Though this difference was statistically significant, it is quite negligible compared to the normalization shown by the network. For the network, the corresponding estimated phoneme boundaries move from 49 ms to 78.5 ms of VOT; by almost 30 ms instead of the 1.5 ms shift seen in the
perception experiments. Evidently this particular neural net, which contains no hidden units and is thus treating the production data as linearly separable, is doing a pretty terrible job of predicting the results of the perception experiment.

Figure 4. Production data (gray and white circles) for VOT in words starting with /g/ or /k/, the phoneme boundaries for the network (NN, black circles) and human listeners (P&P, black triangles).

There were also three stimulus continua involving quantity at the three same different speaking rates, having identical syllable durations as the VOT continua. In these continua durations of vowels and consonants were both varied while the duration of the syllable rhyme was kept constant. Results are shown in Figure 4.

5. CONCLUSIONS

It is clear from Figures 3 and 4 that the network fits the results from the quantity perception experiment much better than the VOT perception experiment. In particular, the slope of the line connecting the network phoneme boundaries for quantity is the same as found in the perception experiments. In the case of VOT the slopes are widely divergent. The reason for this discrepancy can of course be traced to the structure of the networks employed. These contained no hidden units and thus will only solve linearly separable problems. It is interesting that this limitation does not hinder the network from dealing with vowel and consonant durations as cues to quantity and yielding results quite comparable to those found in the perception experiment. In this respect the present simulations reinforce earlier claims [12] that an invariant of V/rhyme can account for the perception of quantity in Icelandic (though see [13] for some limitations). The perception of VOT obviously follows a different path.

REFERENCES