Using artificial neural nets to compare different vocal tract models

Båvegård, M. and Högberg, J.

journal: STL-QPSR
volume: 34
number: 2-3
year: 1993
pages: 007-014

http://www.speech.kth.se/qpsr
USING ARTIFICIAL NEURAL NETS TO COMPARE DIFFERENT VOCAL TRACT MODELS

Mats Båvegård and Jesper Höberg

ABSTRACT

In this study artificial neural nets (ANN), relating articulation to acoustic features, are used as tools to investigate the ability of different vocal-tract models to describe the vocal tract. The vocal tract models are compared by means of a test material common to all evaluations. Different ANN configurations are used and investigated. Aspects of analysing natural speech signal with ANNs, trained on synthetic speech, are discussed. In addition, some effects of the speaker-dependency of the ANNs are investigated. The models are tested on synthetic speech (based on X-ray data of Swedish vowels) and natural speech from different speakers.

INTRODUCTION

Relating articulation to a given acoustic signal is often referred to as the inverse problem. One proposed method is to derive area functions from the speech signal using reflection coefficients obtained by LPC analysis (Wakita, 1979). A second method is to make small perturbations starting from an area-function estimate based on knowledge about the formant frequency relations (Lin, 1990). There are several limitations involved and the results are not always satisfactory.

Another approach is to use artificial neural networks (ANN) as inversion tools. The ANN can provide a reasonable relation between the acoustic signal and the area function without using an exact mathematical transformation.

The performance of ANNs are traditionally evaluated by independent test frames similar to, but not present in, the training material. In this study the training material consists of model-generated area functions and a corresponding acoustic representation.

The test material is based on thirteen area functions of Swedish vowel (Fant, 1992). As test material we also use natural speech of one male and one female speaker.

METHOD

The inversion problem is approached by means of artificial neural networks (ANNs). Three different models are used to generate area functions corresponding to vowel-like articulations. The first is an articulatory model developed by Maeda (1990) using 7 parameters to provide a sagittal description of the vocal tract. Power functions are used to transform the sagittal distances to a cross-sectional area. The next two models provide a direct description of the area function and are controlled by three parameters, the place of constriction, $X_C$, the degree of constriction, $A_C$, and a lip-rounding parameter $l_0/A_0$ where $l_0$ is the length of the lip-section and $A_0$ is the area at the lips.

* This paper was presented at EuroSpeech '93 in Berlin
** Names in alphabetic order
Model II, which originates in the old model by Fant (1960) and was developed by Lin (1990), consists of a straight tube and a tongue hump that is modelled by cosine functions. Model III, Fant (1992), provides less stylized vocal tract shapes that are very similar to area functions estimated from X-ray data of Swedish vowels.

Area functions were generated by varying six out of the seven parameters of Model I independently and the three parameters of Model II and Model III independently.

These area functions were all scaled to the same vocal-tract length, 17.5 cm, and represented as 16 unit length tubes, a constraint imposed by the speech synthesizer.

Speech signals are synthesized from the area functions by a time-domain Line Electric Analog (Liljencrants, 1985) and are low passed using a filter with a cut-off frequency at 5 kHz. The ANNs are trained on the relation between acoustics and articulation. A MEL-scaled filter-bank representation of the speech signal is used as input pattern and the corresponding area function is used as output pattern.

The algorithm used, training the ANNs is the back propagation method (McClelland & Rumelhart, 1988). Different architectures have been tested and each ANN of this study have 16 input nodes, 20 nodes in a hidden layer, and 16 output nodes.

In the first experiment, single ANNs are trained for each vocal-tract model and the resulting nets are tested on the same reference patterns. These reference-area functions are those of 13 Swedish vowels, derived from three dimensional X-ray data, reported in (Fant, 1992). They cover most parts of the Swedish vowel space representing front-, mid- and back vowels. The deviation from the reference-area function is calculated as

\[
\varepsilon = \sum_{i=0}^{15} \left[ \ln A_{\text{ref}}(i) - \ln A(i) - \ln \Delta \right]^2
\]

where \( A_{\text{ref}} \) is the reference area and \( A \) is the area provided by the ANN. The index runs from 0 to 15 since the vocal tract is modelled by 16 tubes. The error \( \varepsilon \) is minimized using a scale factor, \( \Delta \), for the entire area function.

A different network architecture is created by dividing the training material into sub-regions, each region represented by an ANN of its own. Good results have been published by Rahim et al (1993) using an assembly of networks for acoustic to articulatory inversion.

A region corresponds to a rectangular area in the first and second formant plane. That is, all training patterns pertaining to one of the eight regions represent similar articulations. The training material is that generated by Model III. The training materials of each net overlap each other to avoid critical decisions at the borders between neighbouring regions. See Fig. 1.

Two different methods for the detection of the correct region are compared, see Fig. 2. The first is a peak-picking algorithm, tracking the two lowest formant frequencies from the FFT-spectrum. The second approach uses a pre-processing ANN that is trained to associate the filter-bank representation to the correct F1-F2 region.
The multi-ANN architecture is compared to the single ANN using the same synthetic references as in the preceding evaluation but also on natural speech of one male and one female speaker. The two speakers were recorded uttering the same group of vowels that make up the synthetic references to allow for comparisons. The vowels were pronounced in isolation under 'normal conditions', i.e., the recordings were made in an office room and thus some noise is apparent in the signals. This environment was
chosen deliberately since most applications are likely to run in such conditions. No reference area function is available for the two speakers, therefore, the output of the ANN is resynthesized and an additional evaluation criterion is calculated as

\[ \varepsilon = \sum_{i=0}^{13} \left[ fb_{ref}(i) - k \cdot fb(i) \right]^2 \]  

(2)

\( fb_{ref} \) denotes the reference filter bank and \( fb \) is the resynthesized filter-bank value corresponding to the output-area function. The filter-bank coefficients used to calculate equation (2) cover the frequency range 0-5 kHz. The high-frequency register is left out from the evaluation since the energy in this range mostly depends on the low-pass filter used. The error is minimized by optimizing a factor, \( k \), to disregard the absolute level.

A source parameter in the speech-signal synthesis is varied to simulate one aspect of differences between speakers. The voice source is modelled by the LF-model (Fant et al., 1985), a parametric glottal flow model. One of the parameters, \( t_A \), is a time constant describing the return phase from the point of minimum flow derivative. An increase in \( t_A \) implies a decrease in high-frequency energy since the inverse \( t_A \) of this parameter is equivalent to the cut-off frequency of a first-order low-pass filter. Thus we simulate speakers of differing glottal closure characteristics by varying \( t_A \).

RESULTS

The results of testing the ANNs trained on material generated by different vocal-tract models is displayed in Table 1. The figures in the table is the sum of the errors, using equation (1), over all thirteen synthetic vowels. The error is smallest for Model III and greatest for Model II.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \varepsilon(1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>47.7</td>
</tr>
<tr>
<td>Model II</td>
<td>71.6</td>
</tr>
<tr>
<td>Model III</td>
<td>24.6</td>
</tr>
</tbody>
</table>

Table 1. Accumulated error for networks trained on three different vocal tract models using equation (1) for the synthetic test material.

Table 2 indicates that the detection of the F1-F2 region works better with the peak-picking method than with the ANN classifier. However, this pre-processing step is obviously still a weak link since even the better method fails to select the correct net at all instances even though the algorithm was adapted to work for the speakers investigated. It is reasonable to believe that the peak picking method would work less well for other speakers. Further we observe that the error for the multi-net approach, given a correct choice of ANN, is only slightly better than for the single net.
Table 2. Accumulated error calculated using equation (1) for the synthetic test material. PP (peak picking) and ANN refers to the different methods of selecting sub-net. Minimum refers to a 100% correct selection.

<table>
<thead>
<tr>
<th></th>
<th>ε(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single net</td>
<td>24.6</td>
</tr>
<tr>
<td>PP, Multi net</td>
<td>26.5</td>
</tr>
<tr>
<td>ANN, Multi net</td>
<td>73.7</td>
</tr>
<tr>
<td>Minimum, Multi net</td>
<td>22.5</td>
</tr>
</tbody>
</table>

The results for natural speech are displayed in Fig. 3a and 3b, using the deviation in the area function and in the acoustic domain, respectively. It should be emphasized that the area functions of the Swedish vowels used to calculate deviations in Fig. 3b cannot be considered as references. However, we observe the same trend in Fig. 3a and Fig. 3b, the error of the synthetic material does not decrease significantly when introducing a multi-ANN architecture. The deviation decreases for the speech of both the male and the female speaker. Alternative acoustic distance measures using relative formant-frequency errors indicate a greater gain using the multi-ANN approach than the figures presented here.

Figure 3a. The accumulated articulatory deviation calculated using equation (1) for the test material. Black and white bars indicate results of the single-net output and multi-net output, respectively, when the choice of sub-net is correct.

Feeding the ANN with filter-bank values originating from the same area function but synthesized with different values of the source parameter $t_A$ result in slightly different outputs, i.e. different area functions.

The natural speech was recorded at a higher level than the training material. Therefore, the speech samples are also damped to a level corresponding better to that of the training material and the two different versions are compared by tests. These tests indicate that the damped versions provide better outputs suggesting that the absolute level is important.
DISCUSSION AND CONCLUSIONS

Table 1 tells us that Model I11 is the most efficient tool for reproducing the area functions of thirteen Swedish vowels. This is not surprising since Model I11 actually is based on these X-ray derived data. Thus, it is hard to say that one model is better than another in a general sense. However, in some applications (Båvegård & Hörgberg, 1992), where the articulation is displayed visually, it might be of interest to use a model that provides area functions that reflect the actual outlines of the vocal tract. Hence, Model I and Model I11 could be more suitable for such applications.

An acoustic evaluation, using equation (2), also gives at hand that the outputs provided by ANNs trained on Model I and Model II represent poorer results, than those provided by Model I11, in the acoustic domain. That is, the articulations corresponding to the outputs of the different nets are not acoustically equal and thus they do not reflect ambiguity problems that can be encountered when dealing with the inversion problem. At the same time, these results indicate that the training material generated by Model I and Model II do not cover the same acoustic space as the training material provided by Model I11.

We conclude that the ANN is sensitive to the level of the filter-bank representation, both absolute values and the spectral tilt. This seemingly important dependency can probably explain why some inputs have lost their phonemic identity when resynthesized on the output side. For instance, the spectral representation of the male /ø/ does not display an equally steep spectral tilt as its synthetic correspondence. The resynthesized output of the natural vowel has got too low F1 and too high F2, i.e. a vowel with a modest spectral slope.

Ideally, the net should be insensitive to variations in source characteristics since the ANN output model only the supra-glottal tract. However, we have noted that the net indeed is responding to such variations. Thus, some normalization procedure should operate on the input spectral representation to cancel the impact of differing source
characteristics. For instance, the mean spectral slope extracted from a long time average spectrum could be used as a normalizing means to account for the sensitivity discussed above.

One major drawback of the area functions used in the training materials of this investigation is that they were all scaled to the same length without scaling the area dimension. It is well known from acoustic theory that this will affect the spectrum. Evidently it is desirable to have a more flexible system in future work capable of handling speech originating from speakers with different vocal-tract dimensions.

It is also important to note that the training materials of the different models are not of equal size. However, it is not perfectly evident that a larger training set will result in a better acoustic to articulatory mapping (Båvegård & Högb erg, 1992).

ACKNOWLEDGEMENT

This work is in part supported by a grant from the Swedish Language Technology Programme financed by NUTEK and HSFR.

REFERENCES


