Language detection by neural discrimination

Sebastien Herry¹, Bruno Gas², Celestin Sedogbo¹, Jean-Luc Zarader²

¹ Human Interaction Technologies, Advance Software Department
Thales Research & Technology France
sebastien.herry@thalesgroup.com celestin.sedogbo@thalesgroup.com

² Laboratoire des Instruments et Systèmes d’Île de France
Université Pierre et Marie Curie
gas@ccr.jussieu.fr zarader@ccr.jussieu.fr

Abstract
In this paper, we present a new method of language detection. This method is based on language pair discrimination using neural networks as classifier of acoustic features. No acoustic decomposition of the speech signal is needed. We present an application of our method for the detection of 11 languages for a signal duration of 3 seconds (OGI MLTS corpus). The obtained results highlight scores ranging from 75.1% to 82% according to the language processed.

1. Introduction
Language detection is the process which decides if a language is spoken or not in a speech stream. It is a part of the Language Identification (LID) which determines the language spoken from a set of given language (English, Farsi, French, German...).

LID takes benefit of the interest for multilingual system which target the international call center for example. The techniques usually used in LID research are based on spectral parameters distribution modeling and/or language modeling by n-gram for the phoneme series. The most used is the phonotactic approach which represents a first compromise between the level of knowledge required for the learning and the identification accuracy. The phonotactic system observes the phoneme series to establish a statistical model of language, like the model proposed by Zissman [1]. It is based on the Parallel Phone Recognition followed by Language Modeling, which needs Acoustic-Phonetic Decoders (APD).

This imposes a heavy constraint, because APDs need phonetically labeled corpus which are only available for few languages. Techniques have been developed which bypass this problem, for example P. Torres [2] replaces phonemes by automatically created acoustic units. J. Farinas [3] uses also a creation of pseudo-phonemes, and in the past Y. Muthusamy [4] used a technique that generates automatically phonetic macro-class. Unlike the precedents authors, W. Wu & C. Kwasny [5] model the speech signal with recurrent neural networks without acoustic units. They make language identification (English and French, with 4 speakers from audio file of 12.5s duration) by acoustic vector series classification.

We propose in this paper a method using the concept of acoustic vector series classification too. It is only based on acoustic discrimination of language between them, with neural networks. We present an application of detection of 11 languages for phone speech signal of 3 seconds duration. Our approach is different of the one of W. Wu & C. Kwasny [5], because we don’t model the signal in the time, we just classify the acoustic vectors and then make a merging of the results. The main objective of our method is to obtain a technique that allows to reduce the speech duration required for identification in order to give an answer more quickly. As 3 seconds is the shortest duration where some results have been published, we chose this duration for our tests for comparison reasons. The detection is based on acoustic discrimination between languages. We start from the hypothesis that language information for a large part is present in the spectrum. We use several neural networks, each of them discriminates a couple of languages. The result of this discrimination is merged by language to provide first signal detection. We then improve the decision by taking into account the signal of detection of the other languages. The continuation of this paper is cut in five parts, the second part describes the corpus used, the third one explains the method detection for a language, the fourth shows the generalization and the improvement of the method, and finally in the fifth one, we discuss about the future works and next we conclude.

2. The data corpus
We use the CORPUS OGI MLTS [6]. This corpus is composed of 11 languages (English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese). For each of them more than 110 speakers have pronounced several sentences of different duration (3s, 10s, and 60s), for a total of more than 30 hours of speech. We have essentially chosen this corpus because of high speaker number and language number. Our objective is to work on sentences of 3 seconds duration, thus we perform a transformation of sentences. All 11 language sentences have been divided into sentence segments of 3s duration. Next, we distribute these parts among 3 subsets: learning, development for tuning the system and test. The ratio of these subsets are respectively 3/5, 1/5, 1/5. As the OGI corpus contains many hours of speech for each language, each part of the subset is significant. For the distribution, we do not distinguish between the sex, the age, or the birthplace, thus all the speakers are included in each subset.

3. Detection of English
3.1. Principle of detection
Our first goal is to detect the English language among 11 languages. To do that we use neural nets, each of them identifies English language versus another language for example English versus French. Thus we are able to
discriminate English from the remainder of corpus, by merging discrimination coming from all the networks. Acoustic vectors are then processed by the 10 networks, each of them giving as output a discriminant signal for a language pair. These signals, computed for all the acoustic vectors extracted from 3s speech duration are then used to calculate an average. Thus we obtain 10 averages (each of them corresponding to a network discriminating a language pair) which are merged to give a global English detection signal.

3.2. The front end processing

We used the front end RASTA processing [7] to generate acoustic vectors of dimension 24 of relevant spectrum representation. This front end is specially designed to noisy speech which is the case of the corpus OGI MLTS. We use the RASTA processing on 32ms of signal with an overlap of 50%. In the process each set of parameters is normalized.

3.3. Discriminant Neural Nets

At this step we train a simple neural network for each pair of language: (English vs. another one) on RASTA acoustic vectors. Each of the networks are composed of 23 inputs, we have not taken into account the 24th energy coefficient which is not very relevant for language discrimination. These networks are multi-layer perceptron (MLP) with sigmoidal activation functions, 50 hidden cells and two outputs (one for each language). The learning algorithm used was the stochastic back propagation algorithm, to increase the learning speed; (L. Bottou [8]).

3.4. Merging local discrimination for English detection

At this stage of the model, we discriminate only pair of language: (English vs. another one) on RASTA acoustic vectors. During the 2nd step, we sum the network outputs during the time by considering only values which are the most relevant for the final decision. Let \( O_{EN-FA}^{EN-FA}(q) \) being the English output of the network: English versus Farsi (En vs. FA), when presenting the acoustic vector \( q \) as input. Outputs \( O_{EN-FA}^{EN-FA}(q) \) and \( O_{FA-EN}^{FA-EN}(q) \) are relevant for the final decision if:

\[
|O_{EN}^{EN-FA}(q) - O_{FA}^{EN-FA}(q)| > \theta_1
\]

Where \( \theta_1 \) is a threshold determined to induce the best score on the learning corpus. Thus we compute the averages of the English outputs during time like:

\[
A_{EN} = \frac{1}{\beta_{EN}} \sum_q O_{EN-FA}^{EN-FA}(q) \beta(q)
\]

with:

\[
\beta(q) = \begin{cases} 
1 & \text{if } |O_{EN}^{EN-FA}(q) - O_{FA}^{EN-FA}(q)| > \theta_1 \\
0 & \text{otherwise}
\end{cases}
\]

\[
\beta_{EN} \text{ being the normalization factor:}
\]

\[
\beta_{EN} = \sum_q \beta(q)
\]

During the third step, we compute the average for all the previously defined outputs corresponding to the detection of English (\( A_{EN} \)):

\[
GA_{EN} = \frac{1}{10} \sum_i A_{EN}(i)
\]

Where \( A_{EN}(i) \) is the previously defined English output average of network \( i \). The global decision (English detection) is given by:

\[
GA_{EN} > \theta_2 \Rightarrow \text{English Detection} \\
GA_{EN} \leq \theta_2 \Rightarrow \text{no English Detection}
\]

Where \( \theta_2 \) is a threshold determined in the same way as \( \theta_1 \).

4. Experimentation

4.1. Language pair identification

We present in this section the first experimental results obtained on pair language identification with our model. In these experiments we have not considered the final decision (English detection) but only the local decisions made from the network outputs. Table 1 presents the scores of identification obtained on the test corpus for each pair of language. The learning is done on acoustic vectors which represent only a duration of 32ms. This duration is currently
used to detect phoneme. For instance we obtained on the first network (see the EN vs. FA column of table 1) an English identification score of 59.9% (59.9% of the 32ms English frame have been recognized as English language frames) and a Farsi identification score of 63.8%. Thus the average identification score of English/Farsi detector was 61.9%.

Table 1: Scores of language pair identification in percent.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>EN vs. FA</th>
<th>EN vs. FR</th>
<th>EN vs. GE</th>
<th>EN vs. HI</th>
<th>EN vs. JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>61.9</td>
<td>63.8</td>
<td>63.8</td>
<td>62.3</td>
<td>61.6</td>
</tr>
</tbody>
</table>

The results involve that the average identification score is 63.2%. The scores of each network are close to each other: the maximum deviation between all networks is about 5%. These results are encouraging because they were obtained from decisions based on frames of only 32ms duration. They are improved up to 73% by considering sentences of 3s duration and using the whole model defined by eq. (2) and (3) (see next section). Table 1 shows that the learning is homogenous, the standard deviation is only 2% for English and 3.3% for other languages, involving that the learning could be extended to others languages as we will see in section 5.

4.2. Results of English detection with rejection

Our simulations have shown that the difference between the average value (English decision) and the threshold decision $\theta_1$ could be used as a quality factor for the final decision. To do that, we have conducted experiments with classification or rejection of samples. A 3s sentence is rejected if the difference between the average value ($G_{avg}$) and $\theta_1$ is less than a rejection threshold $\theta_2$. We have reported Fig. 2 percentage of rejected sentences, and detection rate of non rejected sentences for different values of $\theta_1$ (varying from 0 to 0.42 corresponding to a very high rejection rate, more than 80%).

Figure 2: Rejection curves of English detection in percent.

This figure shows us the increase of English detection rate with the value of the threshold rejection. Without rejection we have 73% of good detection and with 30% of rejection we detect English at 80%, both on sentences of 3s duration. This increase of the detection rate with rejection implies that the range of the detection values could be interpreted as a certainty.

5. Implementation to the other languages

5.1. Reapplication of process

The goal of this part is to extend the English detection model to other languages, that’s why we reapply the principle, for the 11 languages: we train neural networks for all possible pairs of language. As we have 11 languages, for each of them we can create 10 different networks and we perform the evaluation to determine the threshold $\theta_1$ and $\theta_2$. We present in the table 2 the results of each detection system computed as previously for English. “alarm” means score of valid detection of considered language, “silence” means score of valid non-detection (all other languages) and “detection” is the average detection score of alarm and silence.

Table 2: Scores of language detection in percent.

<table>
<thead>
<tr>
<th>Language</th>
<th>EN</th>
<th>FA</th>
<th>FR</th>
<th>GE</th>
<th>HI</th>
<th>JA</th>
<th>KO</th>
<th>MA</th>
<th>SP</th>
<th>TA</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>alarm</td>
<td>71.6</td>
<td>66.6</td>
<td>76.5</td>
<td>71.0</td>
<td>71.0</td>
<td>70.0</td>
<td>69.8</td>
<td>70.9</td>
<td>71.1</td>
<td>79.7</td>
<td>72.8</td>
</tr>
<tr>
<td>silence</td>
<td>74.5</td>
<td>69.0</td>
<td>75.8</td>
<td>71.7</td>
<td>72.7</td>
<td>72.2</td>
<td>75.3</td>
<td>73.7</td>
<td>74.3</td>
<td>79.6</td>
<td>77.2</td>
</tr>
<tr>
<td>detection</td>
<td>73.1</td>
<td>67.8</td>
<td>76.2</td>
<td>71.4</td>
<td>71.8</td>
<td>71.1</td>
<td>72.5</td>
<td>72.3</td>
<td>72.7</td>
<td>79.7</td>
<td>75.0</td>
</tr>
</tbody>
</table>

The scores presented show that the global average is 73.05% without rejection. The rejections for all the detection systems have the same shape as before (section 4.2): the detection rate is increased with level of rejection. This increase implies that the level of answer for each detector makes sense. Moreover each detector works without interact with the others. Common information could be used to improve de detection rate. Thus we have performed a detection reinforcement in the next section.

5.2. Reinforcement of detection

Given one language detection system, the goal of the reinforcement is to provide to this system the knowledge of the decision of all the other language detection systems to improve its own decision. The reinforcement is done by a new layer of neural net (one network for each language). The outputs of the detection system are fully connected to this new neural net layer (see Fig. 3).

Figure 3: Scheme of reinforcement.
The level 1 corresponds to the reapplication of English detection process to the other languages. Each box, for example “Ge y/n” represents a scheme like Fig. 1 applied to a language. In our case German. The level 2 is the reinforcement method: each box in this level symbolize a neural networks which is connected in input to all previous detection systems. Its output works as a new detection system for one language. In fact the global output of the level 2 has the same meaning as the level 1 global output. According to the table 2, we present in table 3 the corresponding table of the level 2 detections.

Table 3 : Scores of detection after reinforcement in percent.

<table>
<thead>
<tr>
<th>EN</th>
<th>FA</th>
<th>FR</th>
<th>GE</th>
<th>HI</th>
<th>JA</th>
<th>KO</th>
<th>MA</th>
<th>SP</th>
<th>TA</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>alarm</td>
<td>78.5</td>
<td>73.9</td>
<td>79.3</td>
<td>76.5</td>
<td>78.0</td>
<td>74.1</td>
<td>76.4</td>
<td>74.3</td>
<td>76.9</td>
<td>85.1</td>
</tr>
<tr>
<td>silence</td>
<td>77.3</td>
<td>76.4</td>
<td>78.9</td>
<td>76.0</td>
<td>76.4</td>
<td>78.6</td>
<td>75.5</td>
<td>77.7</td>
<td>78.5</td>
<td>90.1</td>
</tr>
<tr>
<td>detection</td>
<td>77.1</td>
<td>75.1</td>
<td>79.1</td>
<td>76.3</td>
<td>77.2</td>
<td>76.4</td>
<td>75.9</td>
<td>76.0</td>
<td>77.7</td>
<td>82.6</td>
</tr>
</tbody>
</table>

The global average score is 77.47% which represents an increase of 4.5% compared to the previous detection system. This increase implies that the different models are not correlated. In their article A. Martin and M. Przybocki [9], give the results of an evaluation organized by NIST, where the best score rate on 3s durations of telephone speech is close to 80% on a detection task among 12 languages. Our test is somehow different, we don’t have the same corpus and the protocol is slightly different. However our corpus is of telephone type, with 11 languages and we evaluate our system with sentences of 3s duration. So we can not directly compare the results, but they have the same order of magnitude.

Figure 4 : Rejection curves of global detection after reinforcement in percent.

The Fig. 4 shows the curves of rejection of global detection after reinforcement. We compute the average of all language rejection curves. These curves have an increase rate which is of the same order than the Fig. 2. We obtain for 30% of rejection a gain of 8% of detection. But the rejection curve is very different, it looks like an exponential curve, showing the end of extension possibilities, because the neural network output values are mainly near from maximal values. 60% of the output values are between 0.9 and the maximum 1.0. The amplitude of the answer could be interpreted, as previously, as a certainty, because the detection rate increases with the rejection rate.

6. Future Work

Our language detection system have been tested on sentences of 3 seconds duration. The next future investigation will be to test the system with sentences shorter and longer than 3s, in order to know how the duration modifies the results. Our second work will be to use a different corpus to compare our results precisely. We have planned to perform our experiment using the corpus CaliFriend [10]. The third step will be the transformation of the detection into identification. Obviously our temporal model must be improved, because we just do an average to merge the detection signals in the time. Other methods could be more efficient.

7. Conclusion

We have proposed a method based on neural networks to detect one language among several others and this for all languages involved. The results obtained with our design allow us to detect language with an average competitive rate: 77% on the OGI multilingual corpus for 3s duration sentences. The modeling technique permits a real time operation on a P4 1.7Ghz. This modeling enables to model any language with a speech corpus without the necessity of a phonetic labeling.

8. References