AUTOMATIC LABELLING OF SPEECH SIGNAL INTO PHONETIC EVENTS

H. Kabré, G. Pérennou and N. Vigouroux
Institut de Recherche en Informatique Toulouse, France

ABSTRACT
In this paper, we give the general principles behind an automatic system, developed at IRIT Lab, and capable of labelling speech signal for phonetic purposes. When using this system, the results are secured on English, French and Swedish corpora demonstrate that the labelling operation becomes completely independent from either language, corpus or speaker. Moreover, this operation requires no manual adaptation or training whatsoever.

1. INTRODUCTION
Automatic labelling of speech corpora is an increasingly important problem, when considering present-day development of recorded speech databases - e.g., the DARPA Project ones. In Europe, within the scope of the SAM ESPRIT Project - involving this kind of databases for multilingual corpora - the question has quickly arisen as to how to adapt these various automatic labellers to different languages, and as to how to process speech material without having to resort either to a manual adaptation or to some kind of language, speaker or corpus training. The latter problem is the one considered, here, as we are presenting SAPHO - the phonetic front-end of our automatic labelling system.

2. SYSTEM STRUCTURE
Our labelling system proceeds in two successive main stages:
- in the course of the first one, the signal is both segmented and labelled into phonetic events (SAPHO component);
- in the second stage, these events become aligned onto a phonetic transcription, supplied beforehand by a phonetician (VERIPHONE component).

In Europe, maximun amplitude is evaluated over each one of the 4ms successive frames; this amplitude has undergone a non-linear smoothing (NLSA parameter) that does preserve major instances of signal discontinuity.

In 1d, mean amplitude (energy) over each 8ms frame is evaluated. These two amplitude values can be directly compared to the initial waveform given in 1b.

In [4], various parameters can be found, necessary for recognition, does not have to be used up.

3) An appropriate, limited choice, among these parameters, will bear upon the minimum subset of parameters, to be used up.

3) An appropriate, limited choice, among these parameters, will bear upon the minimum subset of parameters, to be used up.

The quality of various events, obtained, next, when regrouping frames by labels of equal values, while smoothing off any segment that is too short.

Events are obtained, next, when regrouping frames by labels of equal values, while smoothing off any segment that is too short.

3. EVALUATION
The quality of various events, obtained, can be evaluated thanks to the different kinds of results.

The ones given in Table 2, Fig.3, show quality of automatic segmentation, as this compares to manual labelling, when the latter is obtained over EUROM-p French, English and Swedish corpora, without any kind of either manual adaptation (phase) or learning.

These parameters have to be pre-processed in order to achieve an optimal automatic segmentation.

In Fig.1, for example, three parameter forms can be seen to characterize amplitude.

In 1c, maxumum amplitude is evaluated over each one of the 4ms successive frames; this amplitude has undergone a non-linear smoothing (NLSA parameter) that does preserve major instances of signal discontinuity.

In 1d, mean amplitude (energy) over each 8ms frame is evaluated. These two amplitude values can be directly compared to the initial waveform given in 1b.

Fig.1 - Temporal parameters : 1a) Spectrogram, 1b) waveform, 1c) NLSA normalized amplitude, 1d) 8ms mean amplitude, 1e) NLNZ zero crossing rate.

Fig.2 - Table 1: Phonetic events.

| K | strong syll | S | like r | short S |
| W | weak syll | Z | like fric |
| L | acute voc | F | weak fric |
| U | grave voc | X | x like fric |
| O | voiced occl | Q | unvoiced occl |

Fig.2 - Table 1: Phonetic events.

450

451
These results remain very steady from one language to the next. Furthermore, boundary accuracy is of the same order as what is observed when comparing between 'handlabellers' performances.

<table>
<thead>
<tr>
<th>Pass</th>
<th>Language</th>
<th>Phone Number</th>
<th>Speaker Number</th>
<th>Surfeg Ratio</th>
<th>QC(*)</th>
<th>QC ±17 ms</th>
<th>QC ±21 ms</th>
<th>QC ±25 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>4476</td>
<td>4</td>
<td>2.10</td>
<td>0.850</td>
<td>0.890</td>
<td>0.914</td>
<td>0.933</td>
</tr>
<tr>
<td>2</td>
<td>English</td>
<td>4476</td>
<td>4</td>
<td>1.44</td>
<td>0.794</td>
<td>0.843</td>
<td>0.872</td>
<td>0.888</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>4476</td>
<td>4</td>
<td>1.27</td>
<td>0.763</td>
<td>0.815</td>
<td>0.846</td>
<td>0.865</td>
</tr>
<tr>
<td>1</td>
<td>French</td>
<td>2909</td>
<td>2</td>
<td>2.74</td>
<td>0.866</td>
<td>0.918</td>
<td>0.938</td>
<td>0.950</td>
</tr>
<tr>
<td>2</td>
<td>French</td>
<td>2909</td>
<td>2</td>
<td>1.78</td>
<td>0.823</td>
<td>0.849</td>
<td>0.897</td>
<td>0.933</td>
</tr>
<tr>
<td>3</td>
<td>French</td>
<td>2909</td>
<td>2</td>
<td>1.53</td>
<td>0.939</td>
<td>0.870</td>
<td>0.898</td>
<td>0.912</td>
</tr>
<tr>
<td>1</td>
<td>Swedish</td>
<td>1379</td>
<td>1</td>
<td>2.15</td>
<td>0.782</td>
<td>0.833</td>
<td>0.861</td>
<td>0.884</td>
</tr>
<tr>
<td>2</td>
<td>Swedish</td>
<td>1379</td>
<td>1</td>
<td>1.58</td>
<td>0.734</td>
<td>0.793</td>
<td>0.826</td>
<td>0.846</td>
</tr>
<tr>
<td>3</td>
<td>Swedish</td>
<td>1379</td>
<td>1</td>
<td>1.35</td>
<td>0.700</td>
<td>0.762</td>
<td>0.803</td>
<td>0.822</td>
</tr>
</tbody>
</table>

(*QC = mean/n, N = number of manual boundaries, n = those ones that have as approximation to an automatic boundary less than ±x ms (x=13, 17, 21, 25) Figure 2.3— Table 2: Quality of the SAPHO segmentation.

The other results also display great steadiness, both over various corpora and from one language to the next. The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process. Modeling a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In addition to this process—which is likely to occur in every language—there are properties—also common to all languages—such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process. Modeling a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In addition to this process—which is likely to occur in every language—there are properties—also common to all languages—such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

The results, presented here, were secured with the SAPHO System, which makes use of information relating only to amplitude and zero crossing rate. We are now working at an efficient use of these events in automatic alignment and on pre-selection of sub-vocabularies within large lexicons.

The authors are thankful to Prof. J. F. Malet, CSU Sacramento, for this prompt translation of their original French ms.

5. REFERENCES
[3] S. Benett, V. Zue "Transcription and Alignment of the TIMIT Database", in Getting Started With The DARPA TIMIT CD-ROM.

4. CONCLUSION
The results we have secured over English, French and Swedish speech corpora, demonstrate the feasibility of labelling phonetic events that are language-, speaker-, as well as corpus-independent. However, these results should be reinforced both over larger corpora and over a more numerous set of languages. The results, presented here, were secured with the SAPHO System, which makes use of information relating only to amplitude and zero crossing rate. We are now working at an efficient use of these events in automatic alignment and on pre-selection of sub-vocabularies within large lexicons.

The authors are thankful to Prof. J. F. Malet, CSU Sacramento, for this prompt translation of their original French ms.

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process. Modeling a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In addition to this process—which is likely to occur in every language—there are properties—also common to all languages—such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process. Modeling a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In addition to this process—which is likely to occur in every language—there are properties—also common to all languages—such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process. Modeling a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In addition to this process—which is likely to occur in every language—there are properties—also common to all languages—such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.