Text island spotting in large speech databases
Benjamin Lecouteux, Georges Linarès, Frédéric Beaugendre, Pascal Nocera

To cite this version:
Benjamin Lecouteux, Georges Linarès, Frédéric Beaugendre, Pascal Nocera. Text island spotting in large speech databases. Interspeech, 2007, Anvers, Belgium. hal-02088836

HAL Id: hal-02088836
https://hal.archives-ouvertes.fr/hal-02088836
Submitted on 3 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Text island spotting in large speech databases
Benjamin Lecouteux, Georges Linarès, Frédéric Beaugendre, Pascal Nocéra

This competition is arbitrated by a matching score \( W_i \).

\[ W_i(h_c) = \left| C_i(t) \right| \left| h_c(t) \right| \sum_{i=0}^{n} Idf(w_i) \]

where \( C_i(t) \) and \( h_c(t) \) are the cardinality of respectively the cluster \( C_i \) and the current hypothesis \( h_c \). \( Idf(w) \) represents the classical measure of the relative word frequency:

\[ idf(w) = \frac{1}{\text{frequency}_{w}} \]

Experimental context:
- First experiments assessed on 3 hours of radio ESTER (with exact transcript and a 10% WER transcripts)
- Second experiments assessed on 11 hours of RTBF on which time stamps were manually added.
- All words available in database are added to the language model
- Language model: about 67000 words trained on “l’émancipé”

Results:

<table>
<thead>
<tr>
<th>Radio station</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
<th>Seg. number</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTER</td>
<td>90.9%</td>
<td>98.8%</td>
<td>94.1%</td>
<td>472</td>
</tr>
<tr>
<td>TNFO</td>
<td>93.7%</td>
<td>92.9%</td>
<td>94.9%</td>
<td>468</td>
</tr>
<tr>
<td>RTBF</td>
<td>98.9%</td>
<td>97.8%</td>
<td>98.6%</td>
<td>812</td>
</tr>
<tr>
<td>Mean</td>
<td>95.3%</td>
<td>97.1%</td>
<td>96.1%</td>
<td>1758</td>
</tr>
</tbody>
</table>

Conclusion:
- On ESTER tests approximative transcripts bring a WER gain of about 14% relative, while exact ones allows a WER gain close to 24% relative.
- Spotting performance is good; more than 95.3% of segments have been found, with a precision of about 96.7%.
- On RTBF tests, spotting performance is good; more than 95.3% of segments have been found, with a precision of about 96.7%.

Fast-match to transcript island
- The principle of the proposed method is close to approaches used in the field of information retrieval.
- In our case, the hypothesis is a query which may be answered by one of the transcript islands.
- The lexicon is represented by a lexical space \( L_s \) where each dimension is associated to a word. The coefficients of these vectors represent the frequencies of words in the document.
- As the current hypothesis is developed, a set of word clusters \( C_i \) is built and updated.
- These clusters result from the intersection of \( h_c \) and the transcript island \( I_i \).
- For each new word added to the hypothesis \( h_c \), transcript islands are considered as candidates for guiding the search.