In order to allow content-based search in video and audio files, just as any web search engine does with text [e.g. Google, Yahoo, ...], an indexing of the content has to be produced. The context of this work is the use of automatic speech recognition to develop a tool for semi-automatic indexation of broadcast news.

**Context**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Transcriptions</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection</td>
<td>Acoustic Model</td>
<td>Decoding</td>
</tr>
</tbody>
</table>

Figure 1: Screenshot from the indexation interface

**Keyword Spotting**

2 possibilities:
- Large Vocabulary Continuous Speech Recognition (transcription of the whole)
- Keyword Spotting (Detection of keywords in the continuous speech)

Particularities of the broadcast news speech:
- High variability in speech quality
- Very dynamic vocabulary ➔ selection of keywords in text files of the same day as the audio

Online Garbage system:
Garbage systems attempt to model everything that is not the keyword(s) by one or several garbage model. In online garbage systems, the garbage model is made of a unique garbage phoneme whose probability is computed as the average of the N best states probabilities given the current acoustic vector.

Thresholding:
At a word level, the acoustic score is compared to a threshold. The acoustic score of a word is the accumulation of the acoustic scores of the frames corresponding to this word.

Training of a Support Vector Machine on vectors of a priori features to try to predict if a word is much or less confusing.

The results of classification are used to modify the thresholds of decision on the acoustic scores in order to improve detection results.

**Phonemic content-based threshold adaptation**

A priori phonemic features:
- Transcription length (number of phonemes)
- N-grams of phonemes
- Acoustic confidence score given by the diagonal values of the confusion matrix of the neural network
- Attempt to give a phonemic “eccentricity” value by k-means clustering of symbolic strings (phonetic transcriptions), and measuring distances between a transcription and the centroids of the clusters

ROC curves (>= 4 phones)

ROC curves (>=7 phones)

Figure 3: Bloc diagram of a keyword spotting system

Figure 4: Illustration of a hybrid HMM/ANN system

Figure 5: Illustration of the grammar