An Audio Transcriber for Broadcast Document Indexation

VECSYS
3 r. de la Terre de Feu - Les Ulis
91952 Courtabœuf, France
Contact: Bernard Prouts
bprouts@vecsys.fr

LIMSI - CNRS
B.P. 133
91403 Orsay, France
Contact: Jean-Luc Gauvain
gauvain@limsi.fr

With the rapid expansion of different media sources for information dissemination, there is a pressing need for automatic processing of the audio data stream. For the most part todays methods for segmentation, transcription and indexation are manual, with humans reading, listening and watching, annotating topics and selecting items of interest for the user. Automation of some of these activities can allow more information sources to be covered and significantly reduce processing costs while eliminating tedious work. Some existing applications that could greatly benefit from new technology are the creation and access to digital multimedia libraries (disclosure of the information content and content-based indexation), media monitoring services (selective dissemination of information based on automatic detection of topics of interest) as well as new emerging applications such as News on Demand and Internet watch services. Such applications are now feasible due to the large technological progress in speech recognition made over the last decade, benefiting from advances in micro-electronics which have facilitated the implementation of more complex models and algorithms.

Automatic speech recognition is a key technology for audio and video indexing. Most of the linguistic information is encoded in the audio channel of video data, which once transcribed can be accessed using text-based tools. This is in contrast to the image data for which no common description language is available.

The demonstrator shows state-of-the-art automatic transcription capabilities for indexation of broadcast data in four languages: American English, French, German and Mandarin. Broadcast shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic natures. The main components of the system are the audio partitioner and the speech recognizer, and optionally the topic detector. All are based on statistical modeling techniques. Data partitioning is based on a language independent iterative maximum likelihood segmentation/clustering procedure using Gaussian mixture models and agglomerative clustering. The speech recognizer makes use of continuous density HMMs with Gaussian mixture for acoustic modeling and 4-gram statistics estimated on large text corpora. Word recognition is performed in multiple passes, where initial hypotheses are used for cluster-based acoustic model adaptation to improve word graph generation.

We have recently transcribed about 600 hours of unpartitioned, unrestricted American English broadcast data (TV and radio). The average word error measured on a randomly selected 10 hours subset of this data is 21.5%. The word errors on French and German broadcast news are comparable, about 23%. For Mandarin a character error rate of 20% was obtained. However, not all errors are important for information retrieval as was recently demonstrated in the TREC-8 SDR track. This is particularly true for French where many errors are due to missing agreement of the gender or number of a verb and adjective. Since most information retrieval systems first normalise word forms (stemming) in general these types of errors do not significantly affect information retrieval performance.

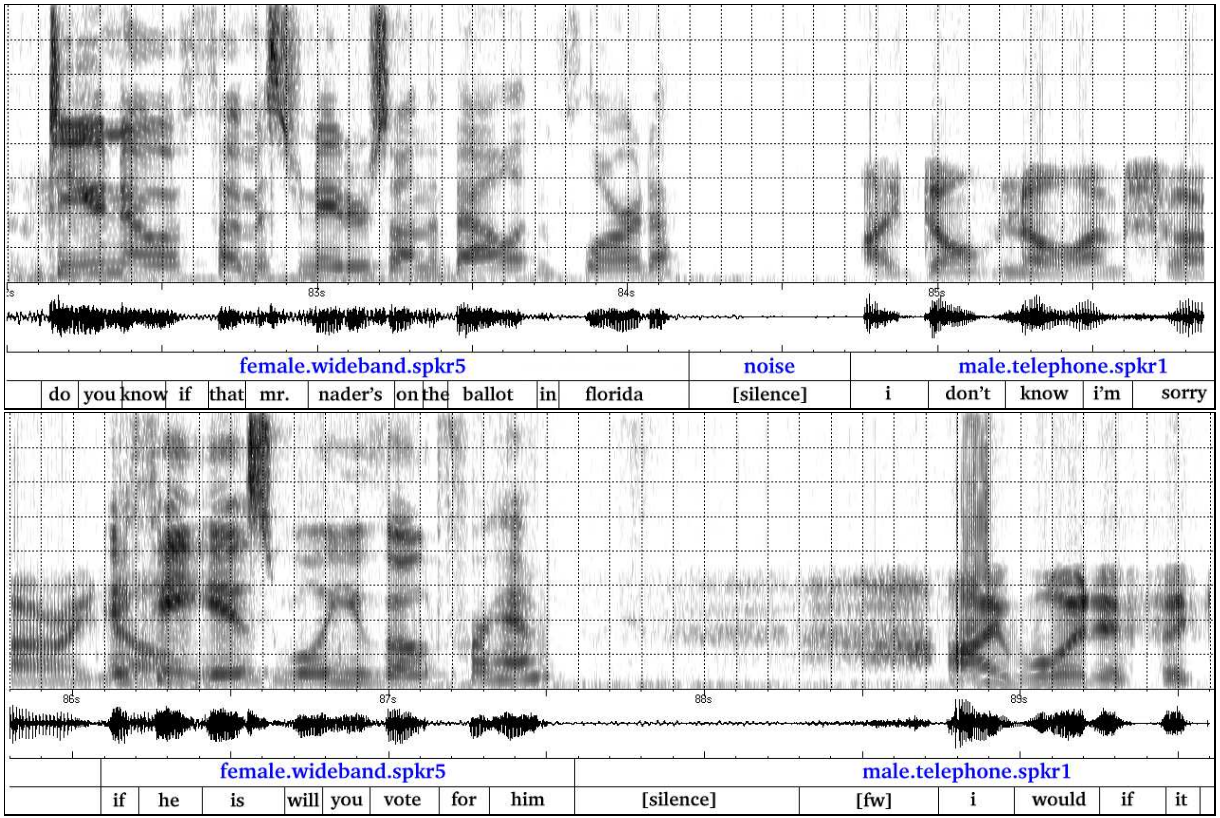
The spoken document retrieval demonstrator returns audio and/or video segments matching a typed natural language query. The extracts are selected from automatically derived transcriptions of shows. The demonstrator also displays the result of the partitioning process (speaker and acoustic condition labels), and the speech transcriptions synchronized with the audio signal.

Reference

Gauvain, J.L., Lamel, L. & Adda, G. (2000). Transcribing broadcast news for audio and video indexing, *Communications of the ACM*, 43(2).

Acknowledgements

This work has been partially funded by the European Commission in the context of the Language Engineering OLIVE project and by the French Ministry of Defense (DGA). The French and German audio and video data have been provided by INA, the French national audio-video archive and ARTE, a broadcasting organization partners in the OLIVE project. The American English and Mandarin data were distributed by the Linguistic Data Consortium.



```
<audiofile  filename=CSPAN-WJ-960917  language=English>
  <segment type=wideband gender=female spkr=5 s_time=81.6 e_time=84.2>
     do you know if that mr. nader's on the ballot in florida
  </segment>
  <segment type=telephone gender=male spkr=1 s_time=84.72 e_time=86.09>
     <w_time s_time=84.72 e_time=84.97> i
    <w_time s_time=84.97 e_time=85.22> don't
    <w_time s_time=85.22 e_time=85.47> know
    <w_time s_time=85.47 e_time=85.63> i'm
    <w_time s_time=85.63 e_time=86.09> sorry
  </segment>
  <segment type=wideband gender=female spkr=5 s_time=86.09 e_time=87.59>
     <w_time s_time=86.09 e_time=86.21> if
    <w_time s_time=86.21 e_time=86.41> he
    <w_time s_time=86.41 e_time=86.67> is
    <w_time s_time=86.67 e_time=86.79> will
    <w_time s_time=86.79 e_time=86.94> you
    <w_time s_time=86.94 e_time=87.16> vote
    <w_time s_time=87.16 e_time=87.32> for
    <w_time s_time=87.32 e_time=87.59> him
  </segment>
  <segment type=telephone gender=male spkr=1 s_time=87.59 e_time=106.22>
     i would if it ...
  </segment>
 </audiofile>
```

Figure 1: Top: Spectrogram illustrating automatic segmentation and transcription output for a segment extracted from a television broadcast. The upper transcript shows the automatically generated partition with labels for segment type: speech (wide-band or telephone), music, or noise; gender; and speaker number. The lower transcription corresponds to the hypothesized word string. Bottom: Example SGML format for the system output. For each segment the signal type, gender and speaker labels, and start and end times are given, as well as the word transcription. For simplicity not all time codes are shown.