Automatically keeping up-to-date with rapidly changing news broadcasts.

Transcribing Broadcast News FOR Audio AND Video Indexing

Jean-Luc Gauvain, Lori Lamel, and Gilles Adda

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, current 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, such as are under exploration in the OLIVE project), 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 feasible due to the large technological progress made over the last decade, benefiting from advances in microelectronics that 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 tran-

scribed can be accessed using text-based tools. This is in contrast to the image data for which no common description language is available. Aithough the focus of this article is on the markup and transcription of the audio channel, some experimental results in spoken document retrieval are also provided.

Background

Radio and television broadcast shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic natures. The signal may be of studio quality or may have been transmitted over a telephone or other noisy channel (that is, corrupted by additive noise and nonlinear distortions), or can contain speech over music or pure music segments.

Gradual transitions between segments occur when there is background music or noise with changing volume, and abrupt changes are common when there is switching between speakers in different locations. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, and so forth. Speech from the same speaker may occur in different parts of the broadcast, and with different background noise conditions. The linguistic style ranges from prepared speech to spontaneous speech. Acoustic and language modeling must accurately account for this varied data.

In the speech recognition community, broadcast news data is often referred to as "found" data, to differentiate it from the type of data common in speech recognition tasks. Up until now speech recognizers have been confronted primarily with read or prepared speech, as in dictation tasks where the speech data is produced with the purpose of being transcribed by the machine, or with limited domain spontaneous speech in more-or-less system-driven dialogue systems. In all cases, users can adapt their language to improve the recognition performance, which can be crucial for some applications. Another interesting aspect of the broadcast news domain is that, at least for what concerns major news events, similar topics are simultaneously covered in different emissions and in different countries and languages. Automatic processing carried out on contemporaneous data sources in different languages can serve for multilingual indexation and retrieval. Multilinguality is thus of particular interest for media watch applications, where news may first break in another country or language.

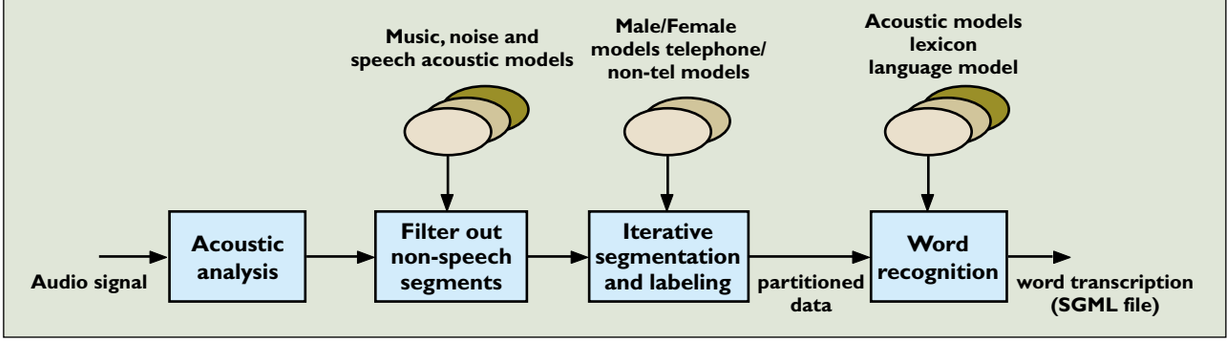
Because of the fast, changing nature of news, one of the main challenges is to keep the models up-to-date. New topics appear suddenly, and remain popular for quite variable length time periods. For example, election coverage may be a hot topic for several months, whereas coverage of natural disasters may last several weeks, and then reappear when a similar event arises. One of the most difficult problems is to quickly be able to recognize previously unseen or rare proper names. Fortunately other sources of contemporary data are available to help keep the system up-to-date, such as written documents from newspapers, newswires, the Internet and subtitles.

At LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) we have been working on using statistical models to transcribe broadcast news data since 1996. Due to the availability of large audio and textual corpora via the Linguistic Data Consortium (LDC),[1] most of our work

[1] www.ldc.upenn.edu

TONY STONE IMAGES

Figure 1. Overview of transcription system for audio stream.



on broadcast news transcription has been carried out on American English. In the context of the European Community (EC) Language Engineering (LE) OLIVE project, broadcast news transcription systems for French and German have recently also been developed. One of the major obstacles in porting across languages is that large corpora of broadcast data are needed to train the statistical models. Obtaining the audio stream is straightforward—just adjust the radio or TV tuner to the desired station. However, this is insufficient as todays statistically based techniques need more supervision, which is provided in the form of accurate transcriptions of the audio data. Other labor intensive steps include building the recognition lexicon—the words that are known to the system, and telling the system how each word is pronounced. Evidently the set of words known to the system is dependent upon the language, as is the set of phonemes used to describe each word. However, it is common to find foreign words in broadcast data particularly for proper names and places. The pronunciation of foreign words can be quite variable depending upon the talker's knowledge of the foreign language. Commercially available transcripts are a good source of training texts, but are often not available in large quantities, and in some countries such transcripts cannot be freely commercialized. These differ from detailed transcriptions in that many spontaneous speech effects (hesitations, word fragments and repetitions) are not transcribed and non-speech events are not labeled. For American English we made use of over 10,000 hours of commercial transcripts available via the LDC. The other data sources such as closed captions and newspaper/newswire texts are more easily accessible but differ substantially from the spoken form. Although such resources may be available on CD-ROM or online for dominant languages, finding large sources of text material may be quite difficult for less prevalent languages.

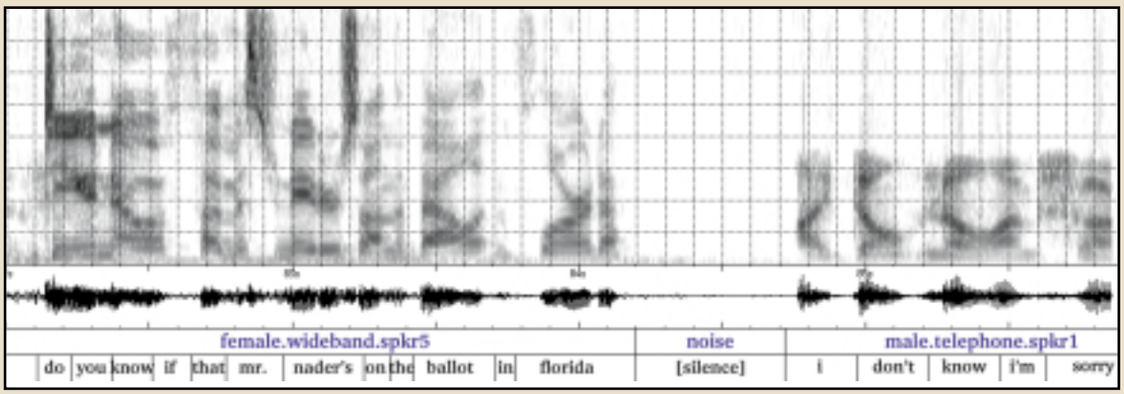
Two principal types of problems are encountered in automatically transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training acoustic models for the different acoustic conditions. Noise compensation is also needed in order to achieve acceptable performance levels. Most broadcast news transcription systems make use of unsupervised acoustic model adaptation as opposed to noise cancellation, which allow adaptation without an explicit noise model. In order to address variability observed in the linguistic properties, the differences in read and spontaneous speech, with regard to lexical items, word and word sequence pronunciations, and the frequencies and distribution of hesitations, filler words, and respiration noises have to be analyzed and modeled in both the acoustic and language models [2].

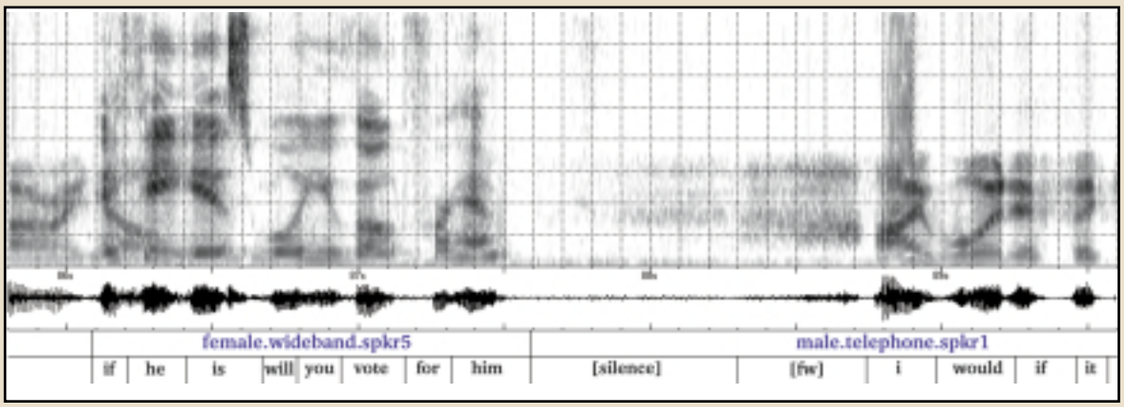
Figure 1 shows the components of a transcription system for broadcast data. The goal of data partitioning (second and third boxes) is to divide the acoustic signal into homogeneous segments, and to associate appropriate labels with each segment. The word recognizer determines the sequence of words in the segment, associating start and end times and a confidence measure with each word.

Partitioning

Prior to word recognition, the data is partitioned into homogeneous acoustic segments. Non-speech segments are identified and removed, and the speech segments are clustered and labeled according to bandwidth and gender. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straightforward solution. First,

Figure 2. Top: Spectrogram illustrating the results of automatic segmentation and transcription on a sequence extracted from a television broadcast. The upper transcript is the automatically generated partition with labels for segment type: speech (wideband or telephone), music, or noise; gender; and speaker number. The lower transcription is 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 (green). For simplicity not all time codes are shown. Transcription errors are shown in red.






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

An essential component of the system is the recognition lexicon, which provides the link between the lexical entries used by the language model and the acoustic models.

in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities. Prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. By using acoustic models trained on particular acoustic conditions (such as wide-band or telephone band), overall performance can be significantly improved. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), substantially reduces the computation time and simplifies decoding (the process of searching the space of all possible word sequences to find the most likely one given the signal and the models).

Data partitioning in the LIMSI transcription system is based on an iterative maximum likelihood segmentation/clustering procedure using Gaussian mixture models and agglomerative clustering. In contrast to partitioning algorithms that incorporate phoneme recognition, this approach is language independent, and the same models are used to partition English, French and German data. The result of the partitioning process is a set of speech segments with speaker, gender and telephone/wideband labels. Two types of segmentation errors can be measured: the extent of the segments and the assigned label. With this approach the average centisecond frame level segmentation error measured on 2 hours of test data is 3.7%. A cluster of segments usually represents a speaker in a given acoustic environment. Thus, there are typically slightly more clusters than true speakers in a show. For example a given speaker's data can be divided into two clusters, one corresponding to speech in presence of background music and the other without music. The cluster purity, defined as the percentage of the audio data in the given cluster associated with the most represented speaker in the cluster is 96% on the same two-hour test set. When clusters are impure, they tend to include speakers in similar noisy background acoustic conditions.

The output of the partitioning process is an SGML file, with one tag per segment specifying the cluster attributes (gender, type and speaker label), and the segment start and end times, shown in blue in the lower part of Figure 2. Based on our experience, it appears that current word recognition performance is not critically dependent upon the partitioning accuracy and that any reasonable approach that marks speaker turns and major acoustic boundaries is sufficient. In fact, many of the partitioning errors occur at the boundary between segments, and can involve silence segments which can be considered as speech or non-speech without influencing transcription performance.

Automatic Transcription

The LIMSI speaker-independent, large vocabulary, continuous speech recognizer makes use of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling. For American English the acoustic models were trained on about 150 hours of broadcast news data distributed by the LDC. The data is taken from a variety of radio and television sources: ABC ("Nightline," "World News Now," "World News Tonight"), CNN ("Early Edition," "Early Prime," "Headline News," "Prime News," "Prime Time Live," "The World Today"), CSPAN ("Washington Journal," "Public Policy"), PRI ("The World") and NPR ("All Things Considered," "Marketplace"). For French the data comes from France Inter (radio) and Antenne-2, TF1 and ARTE (TV), whereas the German data essentially comes from ARTE (ARTE is a French/German television channel). In order to be robust with respect to the varied acoustic conditions, the acoustic models are trained on all data types: clean speech, speech in the presence of background noise or music, or transmitted over noisy channels. Band-limited acoustic models are used with segments labeled as telephone speech.

Language models are used to model regularities in natural language. The most popular methods, such as statistical *n*-gram models, attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. The linguistic model for a given language is obtained by interpolating multiple models trained on data sets with different linguistic properties, For example, our American English language model was trained on about 200 million words of broadcast news transcriptions, 350 million words of North American business newspapers and Associated Press Wordstream texts, and 1.6 million words corresponding to the transcriptions of

the broadcast news acoustic training data. It should be pointed out that it is not enough to simply pour the data into the training module, prior to estimating the language models the texts need to be cleaned to remove typographical errors, and normalized. The texts are also processed to be closer to a spoken language.

An essential component of the transcription system is the recognition lexicon, which provides the link between the lexical entries (usually words) used by the language model and the acoustic models. Lexical design entails selecting the vocabulary items and determining their pronunciation. Each lexical entry is described as a sequence of elementary units, usually phonemes. The American English pronunciations are based on a 48-phone set (3 of them are used to model non-speech events). For French and German, sets of 37 and 49 phonemes are used, respectively. A pronunciation graph is associated with each word so as to allow for alternate pronunciations.

The American English recognition vocabulary contains 65,000 words and has a lexical coverage of over 99% on the November 1998 NIST benchmark test data. It should be noted that lexical coverage is dependent on the language and the type of text normalization used. An optimized 65,000 word lexicon for French has a lexical coverage of about 97.5%, but for German the coverage of an optimized 65,000 word lexicon is only about 95%. The lower lexical coverages are due to the large number of verb forms and number and gender agreement in French and German compared to English, and for German, case declension and compounding. Morphological decomposition is enticing to improve the lexical coverage for a given size lexicon in German.

Word recognition is performed in three steps: initial hypothesis generation; word graph generation; final hypothesis generation. The initial hypothesis is used in unsupervised cluster-based acoustic model adaptation prior to word graph generation. This step, which aims to reduce the mismatch between the models and the data, is crucial for generating accurate word graphs. The first two steps use trigram language models, and the third step uses a 4-gram language model.

Over the last three years we have developed and tested progressively more sophisticated and accurate systems for American English. These systems have consistently achieved top-level performance in NIST benchmark tests [6]. As part of the SDR'99 TREC-8 evaluation we have recently transcribed about 600 hours of unpartitioned, unrestricted American English broadcast data. The average word error measured on a randomly selected 10 hour subset of this data is 21.5%, which implies that the orthographic form for one out of five words is incorrect. However, not all errors are important for information retrieval. 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 normalize word forms (stemming) in general these types of errors should not affect IR performance.

Spoken Document Retrieval

One of the main motivations for automatic processing of the audio channels of broadcast data is to serve as a basis for automatic disclosure and indexation for information retrieval purposes. While in traditional IR tasks, the result is an ordered set of related documents, for spoken document retrieval (SDR) the result is an ordered set of pointers to temporal excerpts [1]. SDR supports random access to relevant portions of audio or video documents, reducing the time needed to locate recordings in large multimedia databases.

The aim of the LE OLIVE project is to develop an archiving and retrieval system for broadcast data to enable efficient access to large multimedia libraries, such as the French INA audiovisual archive. Disclosure of video material plays an important role for the user organizations, but is too costly to carry out manually for all broadcast data. As a result, the vast majority of data is archived with only minimal annotations. The OLIVE consortium is comprised of users, technology providers and integrators.[2] The project is using state-of-the-art speech and natural language processing technologies. The audio stream is automatically partitioned and the speech segments transcribed and time-coded. The transcription is used to generate an index that is linked to the appropriate portions of the audio or video data. OLIVE is also developing tools for users to query the database, as well as cross-lingual access based on off-line machine translation of the archived documents, and online query translation.

As the OLIVE demonstrator is still under development,[3] the performance in spoken document retrieval using LIMSI's state-of-the-art speech recognition technology was assessed using the SDR'98 TREC-7

[2]The OLIVE project is funded by the European Commission under the Telematics Application Programme in the sector Language Engineering. The project (LE4-8364) started in April 1998 and is scheduled to end in 2000. The OLIVE consortium consists of four user organizations: the broadcasters ARTE and TROS; the French national audiovideo archive, INA; and a large service provider for broadcasting and TV productions, NOB. The technology providers are LIMSI-CNRS for speech recognition technology, and TNO (Coordinator), University of Twente and DFKI for natural language processing and information retrieval. The VECSYS and VDA companies are carrying out system integration.
[3]See the OLIVE Web site: twentyone.tpd.tno.nl/olive

data. This data consists of about 100 hours of radio and television broadcasts (1997 LDC Hub4 Broadcast News corpus) and contains about 2800 stories with known boundaries. Retrieval performance was compared using automatically generated transcripts and manually produced transcriptions. These results were obtained using standard IR techniques [7] with query expansion based on parallel blind relevance feedback [5]. Query expansion makes use of additional (parallel) sources text data (preferably from the same epoch) as the audio data to locate terms which co-occur with the terms in the original query so as to enrichen it and to be less sensitive to speech recognition errors. The ordered list of retrieved stories was scored using the TREC-EVAL scoring software and the NIST reference assessments. Using the automatically generated transcripts, a mean average precision [1] of 0.56 was obtained. This is very close to the mean average precision of 0.58 using the manual reference transcripts, even though the average word error on this data is 24%. Recent results obtained on a significantly larger data set (600 hours, 21,700 stories) used in SDR'99 show exactly the same trend that, at least with current IR technology, the limiting factor on performance may not be the transcription accuracy.

Conclusion

Statistical modeling techniques have been successfully applied to the transcription of radio and television broadcasts. This is an exciting research area, in that there are many outstanding issues to be addressed to improve the transcription accuracy on this varied data, and at the same time there are near-term applications which can be successfully built upon this technology, even though it is imperfect.

Our experience is that radio news shows are usually easier to transcribe than television news shows, probably due to the fact that only the audio channel is used to transmit the information, whereas for television the audio stream is supported by visual data. The transcription quality is surprisingly good for speech over background music, however, since the background music level is usually set so that the speaker is easily understood by the listener, this also helps the machine. In contrast, the most difficult portions to transcribe (those for which the error rates are highest) are those containing speech from non-native talkers or overlapping speech such as frequently occurs in interviews or voice-over for translated segments.

Although the basic algorithms for model training, segmentation and decoding can be easily ported across languages, the model accuracy is highly dependent upon the availability of large audio and textual corpora that are needed to estimate the model parameters. These resources are more or less available for some of the dominant languages, but may be inaccessible for less economically or politically important languages. Thus porting to new languages may require a large investment in creating the necessary resources. Multilinguality is of particular interest for media watch applications, where news may first be reported in another country or language.

Keeping the language model and lexicon up to date for breaking news stories is an outstanding challenge for all languages. One solution is to find ways to detect and incorporate new information from temporally simultaneous text sources, such as online newspapers, newswires, and Internet news.

Statistical methods are also being applied to go beyond simple transcription to automatically enhance the output format, to assign and detect topics in the shows and to automatically generate brief summaries by highlighting portions of the transcribed text. C

References

1. Garofolo, J.S., Voorhees, E.M., Auzanne, C.G.P., Stanford, V.M., and Lund, B.A. 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of the 7th Text Retrieval Conference TREC-7*, 1999.
2. Gauvain, J.L., Adda, G., Lamel, L., Adda-Decker, M. Transcribing broadcast news: The LIMSI Nov96 Hub4 system. In *Proceedings of the ARPA Speech Recognition Workshop*, (Feb. 1997), 56–63.
3. Gauvain, J.L., Lamel, L., Adda, G., Jardino, M. Recent advances in transcribing television and radio broadcasts. In *Proceedings of the ESCA EuroSpeech'99*, (Budapest, Sept. 1999).
4. de Jong, F., Gauvain, J.L., den Hartog, J., and Netter, K. OLIVE: Speech based video retrieval. In *Proceedings of CBMI'99* (Oct. 1999).
5. Jourlin, P., Johnson, S.E., Spärck Jones, K., Woodland, P.C. General query expansion techniques for spoken document retrieval. In *Proceedings of SIGIR'99* (Aug. 1999).
6. Pallett, D.S., Fiscus, J.G., Garofolo, J.S., Martin, A., and Przybocki, A.M. 1998 broadcast news benchmark test results. In *Proceedings of the DARPA Broadcast News Workshop* (Herndon, VA, Feb. 1999), 5–12.
7. Robertson, S.E., Walker, S., Jones, S., Hancock-Neaulieu, M.M., and Gatford, M. Okapi at TREC-3. In *Proceedings of TREC-3*, (Washington, D.C., Nov. 1994).

Jean-Luc Gauvain (gauvain@limsi.fr) is a senior researcher and the head of the Spoken Language Processing Group at LIMSI-Centre National de la Recherche Scientifique (CNRS) in Orsay, France; www.limsi.fr/tlp

Lori Lamel (lamel@limsi.fr) is a research scientist with the Spoken Language Processing Group at LIMSI-CNRS in Orsay, France; www.limsi.fr/tlp

Gilles Adda (gadda@limsi.fr) is a research engineer with the Spoken Language Processing Group at LIMSI-CNRS in Orsay, France; www.limsi.fr/tlp

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors gratefully acknowledge the participation of Martine Adda-Decker, Michèle Jardino, Yannick Kercadio, Remi Lejeune, Wolfgang Minker, and Patrick Paroubek to this work.