Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 37507, 11 pages
doi:10.1155/2007/37507

*Research Article*

# A Prototype System for Selective Dissemination of Broadcast News in European Portuguese

**R. Amaral,[1,2,3] H. Meinedo,[1,3] D. Caseiro,[1,3] I. Trancoso,[1,3] and J. Neto[1,3]**

[1] *Instituto Superior Técnico, Universidade Técnica de Lisboa, 1049-001 Lisboa, Portugal*
[2] *Escola Superior de Tecnologia, Instituto Politécnico de Setúbal, 2914-503 Setúbal, Portugal*
[3] *Spoken Language Systems Lab L2F, Institute for Systems and Computer Engineering: Research and Development (INESC-ID), 1000-029 Lisboa, Portugal*

Received 8 September 2006; Accepted 14 April 2007

Recommended by Ebroul Izquierdo

This paper describes ongoing work on selective dissemination of broadcast news. Our pipeline system includes several modules: audio preprocessing, speech recognition, and topic segmentation and indexation. The main goal of this work is to study the impact of earlier errors in the last modules. The impact of audio preprocessing errors is quite small on the speech recognition module, but quite significant in terms of topic segmentation. On the other hand, the impact of speech recognition errors on the topic segmentation and indexation modules is almost negligible. The diagnostic of the errors in these modules is a very important step for the improvement of the prototype of a media watch system described in this paper.



## 1. INTRODUCTION

The goal of this paper is to give a current overview of a prototype system for selective dissemination of broadcast news (BN) in European Portuguese. The system is capable of continuously monitoring a TV channel, and searching inside its news shows for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes, and stores the program. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user, indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories.

Although the development of this system started during the past ALERT European Project, we are continuously trying to improve it, since it integrates several core technologies that are within the most important research areas of our group. The first of these core technologies is audio preprocessing (APP) or speaker diarization which aims at speech/nonspeech classification, speaker segmentation, speaker clustering, and gender, and background conditions classification. The second one is automatic speech recognition (ASR) that converts the segments classified as speech into text. The third core technology is topic segmentation (TS) which splits the broadcast news show into constituent stories. The last technology is topic indexation (TI) which assigns one or multiple topics to each story, according to a thematic thesaurus.

The use of a thematic thesaurus for indexation was requested by RTP (*Rádio Televisão Portuguesa*), the Portuguese Public Broadcast Company, and our former partner in the ALERT Project. This thesaurus follows rules which are generally adopted within EBU (European Broadcast Union) and has been used by RTP since 2002 in its daily manual indexation task. It has a hierarchical structure that covers all possible topics, with 22 thematic areas in the first level, and up to 9 lower levels. In our system, we implemented only 3 levels, which are enough to represent the user profile information that we need to match against the topics produced by the indexation module.

Figure 1 illustrates the pipeline structure of the main processing block of our prototype BN selective dissemination system, integrating the four components, preceded and followed by jingle detection and summarization, respectively. All the components produce information that is stored in an XML (Extendible MarkUp Language) file. At the end, this file

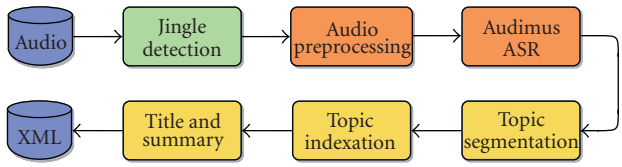


Figure 1: Diagram of the processing block.

contains not only the transcribed text, but also additional information such as the segments duration, the acoustic background classification (e.g., clean/music/noise), the speaker gender, the identification of the speaker cluster, the start and end of each story, and the corresponding topics.

In previous papers [1–3], we have independently described and evaluated each of these components. Here, we will try to give an overview which emphasizes the influence of the performance of the earlier modules on the next ones. This paper is thus structured into four main sections, each one devoted to one of the four modules. Rather than lumping all results together, we will present them individually for each section, in order to be able to better compare the oracle performance of each module with the one in which all previous components are automatic. Before describing each module and the corresponding results, we will describe the corpus that served as the basis for this study. The last section before the conclusions includes a very brief overview of the full prototype system and the results of the field trials that were conducted on it.

A lengthy description of the state of the art of broadcast news systems would be out of the scope of this paper, given the wide range of topics. Joint international evaluation campaigns such as the ones conducted by the National Institute of Standards and Technology (NIST) [4] have been instrumental for the overall progress in this area, but the progress is not the same in all languages. As much as possible, however, we will endeavor to compare our results obtained for a European Portuguese corpus with the state of the art for other languages.

## 2. THE EUROPEAN PORTUGUESE BN CORPUS

The European Portuguese broadcast news corpus, collected in close cooperation with RTP, involves different types of news shows, national and regional, from morning to late evening, including both normal broadcasts and specific ones dedicated to sports and financial news. The corpus is divided into 3 main subsets.

(i) SR (speech recognition): the SR corpus contains around 61 hours of manually transcribed news shows, collected during a period of 3 months, with the primary goal of training acoustic models and adapting the language models of our large vocabulary speech recognition component of our system. The corpus is subdivided into training (51 hours), development (6 hours), and evaluation sets (4 hours). This corpus was also topic labeled manually.

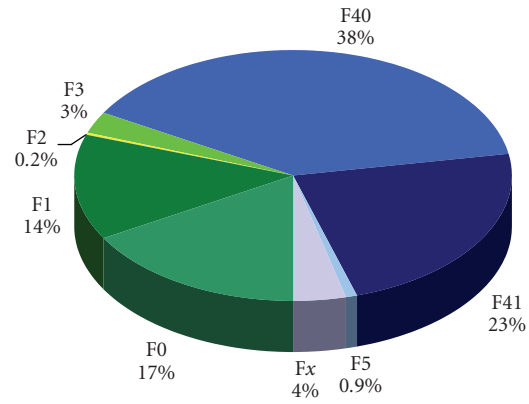


Figure 2: JE focus conditions time distribution: F0 focus condition = planned speech, no background noise, high bandwidth channel, native speech; F1 = spontaneous broadcast speech (clean); F2 = low-fidelity speech (narrowband/telephone); F3 = speech in the presence of background music; F4 = speech under degraded acoustical conditions (F40 = planned; F41 = spontaneous); F5 = nonnative speakers (clean, planned); F$x$ = all other speeches (e.g., spontaneous nonnative).

(ii) TD (topic detection): the TD corpus contains around 300 hours of topic labeled news shows, collected during the following 9 months. All the data were manually segmented into stories or fillers (short segments spoken by the anchor announcing important news that will be reported later), and each story was manually indexed according to the thematic thesaurus. The corresponding orthographic transcriptions were automatically generated by our ASR module.

(iii) JE (joint evaluation): the JE corpus contains around 13 hours, corresponding to the last two weeks of the collection period. It was fully manually transcribed, both in terms of orthographic and topic labels. All the evaluation works described in this paper concern the JE corpus, which justifies describing it in more detail. Figure 2 illustrates the JE contents in terms of focus conditions. Thirty nine percent of its stories are classified using multiple top-level topics.

The JE corpus contains a higher percentage of spontaneous speech (F1 + F41) and a higher percentage of speech under degraded acoustical conditions (F40 + F41) than our SR training corpus.

## 3. AUDIO PREPROCESSING

The APP module (Figure 3) includes five separate components: three for classification (speech/nonspeech, gender, and background), one for speaker clustering and one for acoustic change detection. These components are mostly model-based, making extensive used of feedforward fully connected multilayer perceptrons (MLPs) trained with the backpropagation algorithm on the SR training corpus [1].

The speech/nonspeech module is responsible for identifying audio portions that contain clean speech, and audio

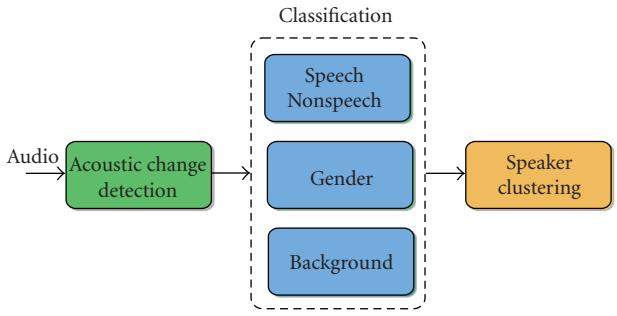


Figure 3: Preprocessing system overview.

portions that instead contain noisy speech or any other sound or noise, such as music, traffic, and so forth. This serves two purposes. First, no time will be wasted trying to recognize audio portions that do not contain speech. Second, it reduces the probability of speaker clustering mistakes.

Gender classification distinguishes between male and female speakers and is used to improve speaker clustering. By clustering separately each gender class, we have a smaller distance matrix when evaluating cluster distances which effectively reduces the search space. It also avoids short segments having opposite gender tags being erroneously clustered together.

Background status classification indicates if the background is clean, has noise, or music. Although it could be used to switch between tuned acoustic models trained separately for each background condition, it is only being used for topic segmentation purposes.

All three classifiers share the same architecture: an MLP with 9 input context frames of 26 coefficients (12th-order perceptual linear prediction (PLP) plus deltas), two hidden layers with 250 sigmoidal units each and the appropriate number of softmax output units (one for each class) which can be viewed as giving a probabilistic estimate of the input frame belonging to that class.

The main goal of the acoustic change detector is to detect audio locations where speakers or background conditions change. When the acoustic change detector hypothesizes the start of a new segment, the first 300 frames of that segment are used to calculate speech/nonspeech, gender, and background classifications. Each classifier computes the decision with the highest average probability over all the frames. This relatively short interval is a tradeoff between performance and the desire for a very low latency time.

The first version of our acoustic change detector used a hybrid two-stage algorithm. The first stage generated a large set of candidate change points which in the second stage were evaluated to eliminate the ones that did not correspond to true speaker change boundaries. The first stage used two complementary algorithms. It started by evaluating, in the cepstral domain, the similarity between two contiguous windows of fixed length that were shifted in time every 10 milliseconds. The evaluation was done using the symmetric Kullback-Liebler distance, KL2 [5], computed over vectors of 12th-order PLP coefficients. This was followed by an energy-based algorithm that detected when the median dropped bellow the long-term average. These two algorithms complemented themselves: energy is good on slow transitions (fade in/out) where KL2 is limited because of the fixed length window. Energy tends to miss the detection of rapid speaker changes for situations with similar energy levels while KL2 does not. The second stage used an MLP classifier, with a large 300-frame input context of acoustic features (12th-order PLP plus log energy) and a hidden layer with 150 sigmoidal units. In practice, the fine tuning of this acoustic change detector version proved too difficult, given the different thresholds one had to optimize.

The current version adopted a much simpler approach: it uses the speech/nonspeech MLP output by additionally smoothing it using a median filter with a window of 0.5 second and thresholding it. Change boundaries are generated for nonspeech segments between 0.5 second and 0.8 second. The 0.8-second value was optimized in the SR training corpus so as to maximize the nonspeech detected.

The goal of speaker clustering is to identify and group together all speech segments that were uttered by the same speaker. After the acoustic change detector signals the existence of a new boundary and the classification modules determine that the new segment contains speech, the first 300 frames of the segment are compared with all the clusters found so far, for the same gender. The segment is merged with the cluster with the lowest distance, provided that it falls bellow a predefined threshold. Twelfth-order PLP plus energy but without deltas was used as feature extraction. The distance measure when comparing two clusters is computed using the Bayesian information criterion (BIC) [6] and can be stated as a model selection criterion where one model is represented by two separated clusters $C_1$ and $C_2$ and the other model represents the clusters joined together $C = \{C1, C2\}$. The BIC expression is given by

$$\mathrm{BIC} = n \log |\Sigma| - n_1 \log |\Sigma_1| - n_2 \log |\Sigma_2| - \lambda \alpha P, \quad (1)$$

where $n = n_1 + n_2$ gives the data size, $\Sigma$ is the covariance matrix, $P$ is a penalty factor related with the number of parameters in the model, and $\lambda$ and $\alpha$ are two thresholds. If BIC $< 0$, the two clusters are joined together. The second threshold $\alpha$ is a cluster adjacency term which favors clustering together consecutive speech segments. Empirically, if the speech segment and the cluster being compared are adjacent (closer in time), the probability of belonging to the same speaker must be higher. The thresholds were tuned in the SR training corpus in order to minimize the diarization error rate (DER) ($\lambda = 2.25$, $\alpha = 1.40$).

### 3.1. Audio preprocessing results

Table 1 summarizes the results for the components of the APP module computed over the JE corpus. Speech/ Nonspeech, gender, and background classification results are reported in terms of percentage of correctly classified frames for each class and accuracy, defined as the ratio between the number of correctly classified frames and the total number

Table 1: Audio preprocessing evaluation results.

| | | | | |
|---|---|---|---|---|
| Speech/nonspeech | | Speech | Nonspeech | Accuracy |
| | | 97.9 | 89.1 | 97.2 |
| Gender | | Male | Female | Accuracy |
| | | 97.4 | 97.8 | 97.5 |
| Background | Clean | Music | Noise | Accuracy |
| | 78.0 | 65.8 | 88.9 | 84.7 |
| Clustering | | Q | Q-map | DER |
| | | 76.2 | 84.4 | 26.1 |

of frames. In order to evaluate the clustering, a bidirectional one-to-one mapping of reference speakers to clusters was computed (NIST rich text transcription evaluation script). The Q-measure is defined as the geometrical mean of the percentage of cluster frames belonging to the correct speaker and the percentage of speaker frames labeled with the correct cluster. Another performance measure is the DER which is computed as the percentage of frames with an incorrect cluster-speaker correspondence.

Besides having evaluated the APP module on the JE corpus, which is very relevant for the following modules, we have also evaluated it on a multilingual BN corpus collected within the framework of a European collaborative action (COST 278—Spoken Language Interaction in Telecommunication). Our APP module was compared against the best algorithms evaluated in [7], having achieved similar results in terms of speech/nonspeech detection and gender classifications. Clustering results were a little worse than the best ones achieved with this corpus (23%), but none of the other approaches use the low latency constraints we are aiming at.

The comparison with other APP results reported in the literature is not so fair, given that the results are obtained with different corpora. In terms of speech/nonspeech detection, performances are quoted around 97% [8], and in terms of gender classification around 98% [8], so our results are very close to the state of the art.

Background conditions classification besides being a rather difficult task is not commonly found in current state of the art audio diarization systems. Nevertheless, our accuracy is still low, which can be partly attributed to the fact that our training and test corpora show much inconsistency in terms of background conditions labeling by the human annotators.

In terms of diarization, better results (below 20%) are reported for agglomerative clustering approaches [8]. This type of offline processing can effectively perform a global optimization in the search space and will be less prone to errors when joining together short speech segments than the online clustering approach we have adopted. This approach not only is doing a local optimization of the search space, but also the low latency constraint involves comparing a very short speech segment with the clusters found so far.

The best speaker clustering systems evaluated in BN tasks achieve DER results around 10% by making use of state-of-the-art speaker identification techniques like feature warping and model adaptation [9]. Such results, however, are reported for BN shows which typically have less than 30 speakers, whereas the BN shows included in the JE corpus have around 80. Nevertheless, we are currently trying to improve our clustering algorithm which still produces a higher number of clusters per speaker.

## 4. AUTOMATIC SPEECH RECOGNITION

The second module in our pipeline system is a hybrid automatic speech recognizer [10] that combines the temporal modeling capabilities of hidden Markov models (HMMs) with the pattern discriminative classification capabilities of MLPs. The acoustic modeling combines phone probabilities generated by several MLPs trained on distinct feature sets: PLP (perceptual linear prediction), Log-RASTA (log-RelAtive SpecTrAl), and MSG (Modulation SpectroGram). Each MLP classifier incorporates local acoustic context via an input window of 13 frames. The resulting network has two nonlinear hidden layers with 1500 units each and 40 softmax output units (38 phones plus silence and breath noises). The vocabulary includes around 57 k words. The lexicon includes multiple pronunciations, totaling 65 k entries. The corresponding out-of-vocabulary (OOV) rate is 1.4%. The language model which is a 4-gram backoff model was created by interpolating a 4-gram newspaper text language model built from over 604 M words with a 3-gram model based on the transcriptions of the SR training set with 532 k words. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in a development set is 112.9.

Our decoder is based on the weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [11]. In this approach, the search space is a large WFST that maps HMMs (or in some cases, observations) to words. This WFST is built by composing various components of the systems represented as WFSTs. In our case, the search space integrates the HMM/MLP topology transducer, the lexicon transducer, and the language model one. Traditionally, this composition and subsequent optimization are done in an offline compilation step. A unique characteristic of our decoder is its ability to compose and optimize the various components of the system in runtime. A specialized WFST composition algorithm was developed [12] that composes and optimizes the lexicon and language model components in a single step. Furthermore, the algorithm can support lazy implementations so that only the fragment of the search space required in runtime is computed. This algorithm is able to perform true composition and determinization of the search space while approximating other operations such as pushing and minimization. This dynamic approach has several advantages relative to the static approach. The first one is memory efficiency, the specialized algorithm requires less memory than the explicit determination algorithm used in the offline compilation step, moreover, since only a small fraction of the search space is computed, it also requires less runtime memory. This memory efficiency allows us to use large 4-gram language models in a single pass of the decoder. Other approaches are forced to use a smaller language model in the first pass and rescore with a larger language model.

Table 2: APP impact on speech recognition.

| Segment boundary | WER % | |
|---|---|---|
| | F0 | All |
| Manual segment boundaries | 11.3 | 23.5 |
| Automatic segment boundaries | 11.5 | 24.0 |

The second advantage is flexibility, the dynamic approach allows for quick runtime reconfiguration of the decoder since the original components are available in runtime and can be quickly adapted or replaced.

### 4.1. Confidence measures

Associating confidence scores to the recognized text is essential for evaluating the impact of potential recognition errors. Hence, confidence scoring was recently integrated in the ASR module. In a first step, the decoder is used to generate the best word and phone sequences, including information about the word and phone boundaries, as well as search space statistics. Then, for each recognized phone, a set of confidence features are extracted from the utterance and from the statistics collected during decoding. The phone confidence features is combined into word-level confidence features. Finally, a maximum entropy classifier is used to classify words as correct or incorrect. The word-level confidence feature set includes various recognition scores (recognition score, acoustic score and word posterior probability [13]), search space statistics, (number of competing hypotheses and number of competing phones), and phone log-likelihood ratios between the hypothesized phone and the best competing one. All features are scaled to the $[0, 1]$ interval. The maximum entropy classifier [14] combines these features according to

$$P(\text{correct} \mid w_i) = \frac{1}{Z(w_i)} \exp\left[\sum_{i=i}^{F} \lambda_i f_i(w_i)\right], \qquad (2)$$

where $w_i$ is the word, $F$ is the number of features, $f_i(w_i)$ is a feature, $Z(w_i)$ is a normalization factor, and $\lambda_i$'s are the model parameters. The detector was trained on the SR training corpus. When evaluated on the JE corpus, an equal error rate of 24% was obtained.

### 4.2. ASR results with manual and automatic preprocessing

Table 2 presents the word error rate (WER) results on the JE corpus, for two different focus conditions (F0 and all conditions), and in two different experiments: according to the manual preprocessing (reference classifications and boundaries) and according to the automatic preprocessing defined by the APP module.

The performance is comparable in both experiments with only 0.5% absolute increase in WER. This increase can be explained by speech/nonspeech classification errors, that is, word deletions caused by noisy speech segments tagged by the auto APP as nonspeech and word insertions caused by noisy nonspeech segments marked by the auto APP as containing speech. The other source for errors is related to different sentence-like units ("semantic," "syntactic," or "sentence" units—SUs) between the manual and the auto APP. Since the auto APP tends to create larger than "real" SUs, the problem seems to be in the language model which is introducing erroneous words (mostly function words) trying to connect different sentences.

In terms of speech recognition, for English, recent systems have performances for word error rate in all conditions less than 16% with real-time (RT) performance [15], and less than 13% with 10 xRT performance [16]. For French, a romance language much closer to Portuguese, the results obtained in the ESTER phase II campaign [17] show a WER for all conditions of 11.9%, and around 10% for clean speech (studio or telephone), to be compared with 17.9% in the presence of background music or noise. This means that the ESTER test data has a much higher percentage of clean conditions. A real-time version of this system obtained 16.8% WER overall in the same ESTER test set. Comparatively, our system which works in real time has 24% WER in the JE corpus which has a large percentage of difficult conditions like speech with background noise.

These results motivate a qualitative analysis of the different types of errors.

(i) Errors due to severe vowel reduction: vowel reduction, including quality change, devoicing, and deletion, is specially important for European Portuguese, being one of the features that distinguishes it from Brazilian Portuguese and that makes it more difficult to learn for a foreign speaker. It may take the form of (1) intraword vowel devoicing; (2) voicing assimilation; and (3) vowel and consonant deletion and coalescence. Both (2) and (3) may occur within and across word boundaries. Contractions are very common, with both partial or full syllable truncation and vowel coalescence. As a result of vowel deletion, rather complex consonant clusters can be formed across word boundaries. Even simple cases, such as the coalescence of the two plosives (e.g., *que conhecem*, "who know"), raise interesting problems of whether they may be adequately modeled by a single acoustic model for the plosive. This type of error is strongly affected by factors such as high speech rate. The relatively high deletion rate may be partly attributed to severe vowel reduction and affects mostly (typically short) function words.

(ii) Errors due to OOVs: this affects namely foreign names. It is known that one OOV term can lead to between 1.6 and 2 additional errors [18].

(iii) Errors in inflected forms: this affects mostly verbal forms (Portuguese verbs typically have above 50 different forms, excluding clitics), and gender and number distinctions in names and adjectives. It is worth exploring the possibility of using some postprocessing parsing step for detecting and hopefully correcting some of these agreement errors. Some of these errors are due to the fact that the correct inflected forms are not included in the lexicon.

(iv) Errors around speech disfluencies: this is the type of error that is most specific of the spontaneous speech, a condition that is fairly frequent in the JE corpus. The frequency of

repetitions, repairs, restarts, and filled pauses is very high in these conditions, in agreement with values of one disfluency every 20 words cited in [19]. Unfortunately, the training corpus for broadcast news included a very small representation of such examples.

(v) Errors due to inconsistent spelling of the manual transcriptions: the most common inconsistencies occur for foreign names or consist of writing the same entries both as separate words and as a single word.

## 5. TOPIC SEGMENTATION

The goal of TS module is to split the broadcast news show into the constituent stories. This may be done taking into account the characteristic structure of broadcast news shows [20]. They typically consist of a sequence of segments that can either be stories or fillers. The fact that all stories start with a segment spoken by the anchor, and are typically further developed by out-of-studio reports and/or interviews is the most important heuristic that can be exploited in this context. Hence, the simplest TS algorithm is the one that starts by defining potential story boundaries in every nonanchor/anchor transition. Other heuristics are obviously necessary. For instance, one must eliminate stories that are too short, because of the difficulty of assigning a topic with so little transcribed material. In these cases, the short story segment is merged with the following one with the same speaker and background. Other nonanchor/anchor transitions are also discarded as story boundaries: the boundaries that correspond to an anchor segment that is too short for a story introduction (even if followed by a long segment from another speaker), and the ones that correspond to an anchor turn inside an interview with multiple turns.

This type of heuristics still fails when all the story is spoken by the anchor, without further reports or interviews, leading to a merge with the next story. In order to avoid this, potential story boundaries are considered in every transition of a nonspeech segment to an anchor segment. More recently, the problem of a thematic anchor (i.e., sports anchor) was also addressed.

The identification of the anchor is done on the basis of the speaker clustering information, as the cluster with the largest number of turns. A minor refinement was recently introduced to account for the cases where there are two anchors (although not present in the JE corpus).

### 5.1. Topic segmentation results with manual and automatic prior processing

The evaluation of the topic segmentation was done using the standard measures recall (% of detected boundaries), precision (% of marks which are genuine boundaries), and F-measure (defined as $2RP/(R+P)$). Table 3 shows the TS results. These results together with the field trials we have conducted [3] show that boundary deletion is a critical problem. In fact, our TS algorithm has several pitfalls: (i) it fails when all the story is spoken by the anchor, without further reports or interviews, and is not followed by a short pause, leading to a merge with the next story; (ii) it fails when the filler is not detected by a speaker/background condition change, and is not followed by a short pause either, also leading to a merge with the next story (19% of the program events are fillers); (iii) it fails when the anchor(s) is/are not correctly identified.

Table 3: Topic segmentation results.

| APP | ASR | Recall % | Precision % | F-measure |
|---|---|---|---|---|
| Manual | Manual | 88.8 | 56.9 | 0.69 |
| Manual | Auto | 88.8 | 54.6 | 0.67 |
| Auto | Auto | 83.2 | 57.2 | 0.68 |

The comparison of the results of the TS module with the state of the art is complicated by the different definitions of *topic*. The major contributions to this area come from two evaluation programs: topic detection and tracking (TDT) and TREC video retrieval (TRECVID), where TREC stands for The Text REtrieval Conference (TREC), both cosponsored by NIST and the US Department of Defense. The TDT evaluation program started in 1999. The tasks under evaluation were the segmentation of the broadcast news stream data from an audio news source into the constituent stories (story segmentation task); to tag incoming stories with topics known by the system (topic tracking task); and to detect and track topics not previously known to the system (topic detection task). The topic notion was defined as "a seminal event or activity, along with all directly related events and activities." Reference [1] as an example, a story about the victims and the damages of a volcanic eruption, will be considered to be a story of the volcanic eruption. This topic definition sets TDT apart from other topic-oriented research that deals with categories of information [2]. In TDT2001, no one submitted results for the segmentation task and, since then, this task was left out from the evaluation programs including the last one, TDT2004.

In 2001 and 2002, the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. This track became an independent evaluation (TRECVID) [3] in 2003. One of the four TRECVID tasks, in the first two campaigns, was devoted to story segmentation on BN programs. Although the TRECVID task used the same story definition adopted in the TDT story segmentation track, there are major differences. TDT was modeled as an online task, whereas TRECVID examines story segmentation in an archival setting, allowing the use of global offline information. Another difference is the fact that in the TRECVID task, the video stream is available to enhance story segmentation. The archival framework of the TRECVID segmentation task is more similar to the segmentation performed in this work. A close look at the best results achieved in TRECVID story segmentation task (F = 0.7) [4] shows our good results, specially considering the lack of video information in our approach.

Table 4: Topic indexation results.

| APP | ASR | Correctness % | Accuracy % |
|---|---|---|---|
| Manual | Manual | 91.5 | 91.3 |
| Manual | Auto w/o conf. | 94.4 | 90.8 |
| Manual | Auto w/conf. | 94.9 | 91.7 |
| Auto | Auto w/conf. | 94.8 | 91.4 |

## 6. TOPIC INDEXATION

Topic identification is a two-stage process that starts with the detection of the most probable top-level story topics and then finds for those topics all the second- and third-level descriptors that are relevant for the indexation.

For each of the 22 top-level domains, topic and nontopic unigram language models were created using the stories of the TD corpus which were preprocessed in order to remove function words and lemmatize the remaining ones. Topic detection is based on the log-likelihood ratio between the topic likelihood $p(W/T_i)$ and the nontopic likelihood $p(W/\overline{T_i})$. The detection of any topic in a story occurs every time the correspondent score is higher than a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics.

In the second step, we count the number of occurrences of the words corresponding to the domain tree leafs and normalize these values with the number of words in the story text. Once the tree leaf occurrences are counted, we go up the tree accumulating in each node all the normalized occurrences from the nodes below [21]. The decision of whether a node concept is relevant for the story is made only at the second and third upper node levels, by comparing the accumulated occurrences with a predefined threshold.

### 6.1. Topic indexation results with manual and automatic prior processing

In order to conduct the topic indexation experiments, we started by choosing the best threshold for the word confidence measure as well as for the topic confidence measure. The tuning of these thresholds was done with the development corpus in the following manner: the word confidence threshold was ranged from 0 to 1, and topic models were created using the correspondent topic material available. Obviously, higher threshold values decrease the amount of automatic transcriptions available to train each topic. Topic indexation was then performed in the development corpus in order to find the topic thresholds corresponding to the best topic accuracy (91.9%). The use of these confidence measures led to rejecting 42% of the original topic training material.

Once the word and topic confidence thresholds were defined, the evaluation of the indexation performance was done for all the stories of the JE corpus, ignoring filler segments. The correctness and accuracy scores obtained using only the top-level topic are shown in Table 4, assuming manually segmented stories. Topic accuracy is defined as the ratio between the number of correct detections minus false detections (false alarms) and the total number of topics. Topic correctness is defined as the ratio between the number of correct detections and the total number of topics. The results for lower levels are very dependent on the amount of training material in each of these lower-level topics (the second level includes over 1600 topic descriptors, and hence very few materials for some topics).

When using topic models created with the nonrejected keywords, we observed a slight decrease in the number of misses and an increase in the number of false alarms. We also observed a slight decrease with manual transcriptions, which we attributed to the fact that the topic models were built using ASR transcriptions.

These results represent a significant improvement over previous versions [2], mainly attributed to allowing multiple topics per story, just as in the manual classification. A close inspection of the table shows similar results for the topic indexation with auto or manual APP. The adoption of the word confidence measure made a small improvement in the indexation results, mainly due to the reduced amount of data to train the topic models. The results are shown in terms of topic classification and not story classification.

The topic indexation task has no parallelism in the state of the art, because it is thesaurus-oriented, using a specific categorization scheme. This type of indexation makes our system significantly different from the ones developed by the French [22] and German [23] partners in the ALERT Project, and from the type of work involved in the TREC spoken document retrieval track [24].

## 7. PROTOTYPE DESCRIPTION

As explained above, the four modules are part of the central *PROCESSING* block of our prototype system for selective dissemination of broadcast news. This central *PROCESSING* block is surrounded by two others: the *CAPTURE* block, responsible for the capture of each of the programs defined to be monitored, and the *SERVICE* block, responsible for the user and database management interface (Figure 4). A simple scheme of semaphores is used to control the overall process [25].

In the *CAPTURE* block, using as input the list of news shows to be monitored, a web script schedules the recordings by downloading from the TV station web site their daily time schedule (expected starting and ending time). Since the actual news show duration is frequently longer than the original schedule, the recording starts 1 minute before and ends 20 minutes later.

The capture script records the specified news show at the defined time using a TV capture board (Pinnacle PCTV Pro) that has direct access to a TV cable network. The recording produces two independent streams: an MPEG-2 video stream and an uncompressed, 44.1 kHz, mono, 16-bit audio stream. When the recording ends, the audio stream is downsampled to 16 kHz, and a flag is generated to trigger the *PROCESSING* block.

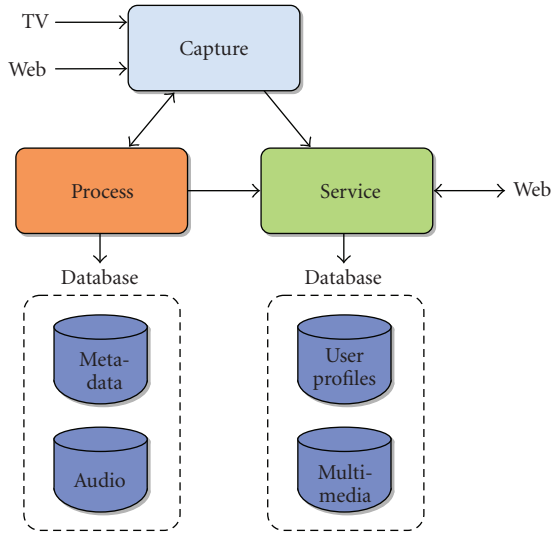


Figure 4: Diagram of the processing block.

When the *PROCESSING* block sends back jingle detection information, the *CAPTURE* block starts multiplexing the recorded video and streams together, cutting out unwanted portions, effectively producing an AVI file with only the news show. This multiplexed AVI file has MPEG-4 video and MP3 audio.

When the *PROCESSING* block finishes, sending back the XML file, the CAPTURE block generates individual AVI video files for each news story identified in this file. These individual AVI files have less video quality which is suitable for streaming to portable devices.

All the AVI video files generated are sent to the *SERVICE* block for conversion to real media format, the format we use for video streaming over the web.

In the *PROCESSING* block, The audio stream generated is processed through several stages that successively segment, transcribe, and index it, as described in preceding sections, compiling the resulting information into an XML file. Although a last stage of summarization is planned, the current version produces a short summary based on the first sentences of the story. This basic extractive summarization technique is relatively effective for broadcast news.

The *SERVICE* block is responsible for loading the XML file into the BN database, converting the AVI video files into real media format (the format we use for video streaming over the web), running the web video streaming server, running the web pages server for the user interface, managing the user profiles in the user database, and sending email alert messages to the users resulting from the match between the news show information and the user profiles.

On the user interface, there is the possibility to sign up for the service, which enables the user to receive alerts on future programs, or to search on the current set of programs for a specific topic. When signing up for the service, the user is asked to define his/her profile. The profile definition is based on a thematic indexation with three hierarchical levels, just as used in the TS module. Additionally, a user can further restrict his/her profile definition to the existence of onomastic and geographical information or a free text string. The profile definition results from an AND logic operator on these four kinds of information.

A user can simultaneously select a set of topics, by multiple selections in a specific thematic level, or by entering different individual topics. The combination of these topics can be done through an "AND" or an "OR" boolean operator.

The alert email messages include information on the name, date, and time of the news broadcast show, a short summary, a URL where one could find the corresponding RealVideo stream, the list of the chosen topic categories that were matched in the story, and a percentage score indicating how well the story matched these categories.

The system has been implemented on a network of 2 normal PCs running Windows and/or Linux. In one of the machines is running the capture and service software and on the other the processing software. The present implementation of the system is focused on demonstrating the usage and features of this system for the 8 o'clock evening news broadcasted by RTP. The system could be scaled according to the set of programs required and the requirement time.

In order to generalize the system to be accessible through portable media, as PDAs or mobile phones, we created a web server system that it is accessible from these mobile devices where the users can check for new stories according to their profile, or search for specific stories. The system uses the same database interface as the normal system with a set of additional features as voice navigation and voice queries.

In order to further explore the system, we are currently working with RTP to improve their website (http://www.rtp.pt) through which a set of programs is available to the public. Although our system currently only provides metadata for the 8 o'clock evening news, it can be easily extended to other broadcast news programs. Through a website, we have all the facilities of streaming video for different kinds of devices and the availability of metadata is starting to be admissible in most of the streaming software of these devices. These communication schemes work on both download and upload with the possibility of querying only the necessary information, television, radio, and text, both in terms of a single program or part of it as specific news.

### 7.1. Field trials

The system was subject to field trials by a small group of users that filled a global evaluation form about the user interface, and one form for each story they had seen in the news show that corresponded to their profile. This form enabled us to compute the percentage of hits (65%) or false alarms (2%), and whether the story boundaries for hits were more or less acceptable, on a 5-level scale (60% of the assigned boundaries were correct, 29% acceptable, and 11% not acceptable).

These results are worse than the ones obtained in the recent evaluation, which we can partly attribute to the improvements that have been done since then (namely, in terms of allowing multiple topics per story), and partly due to the

fact that the JE corpus did not significantly differ in time from the training and development corpora, having adequate lexical and language models, whereas the field trials took place almost two years after when this was no longer true. The continuous adaptation of these models is indeed the topic of an ongoing Ph.D. thesis [26].

Conducting the field trials during a major worldwide event, such as war, had also a great impact on the performance, in terms of the duration of the news show, which may exceed the normal recording times, and namely in terms of the very large percentage of the broadcast that is devoted to this topic. Rather than being classified as a single story, it is typically subdivided into multiple stories on the different aspects of the war at national and international levels, which shows the difficulty of achieving a good balance between grouping under large topics or subdividing into smaller ones.

The field trials also allowed us to evaluate the user interface. One of the most relevant aspects of this interface concerned the user profile definition. As explained above, this profile could involve both free strings and thematic domains or subdomains. As expected, free string matching is more prone to speech recognition errors, specially when involving only a single word that may be erroneously recognized instead of another. Onomastic and geographic classification, for the same reason, is also currently error prone. Although we are currently working on named entity extraction, the current version is based on simple word matching. Thematic matching is more robust in this sense. However, the thesaurus classification using only the top levels is not self-evident for the untrained user. For instance, a significant number of users did not know in which of the 22 top levels a story about an earthquake should be classified.

Notification delay was not an aspect evaluated during the field trials. As explained above, our pipeline processing implied that the processing block only became active after the capture block finished, and the service block only became active after the processing block finished. However, the modification of this alert system to allow parallel processing is relatively easy. In fact, as our recognition system is currently being deployed at RTP for automatic captioning, most of this modification work has already been done and the notification delay may become almost negligible.

On the whole, we found out that having a fully operational system is a must for being able to address user needs in the future in this type of service. Our small panel of potential users was unanimous in finding such type of system very interesting and useful, specially since they were often too busy to watch the full broadcast and with such a service they had the opportunity of watching only the most interesting parts. In spite of the frequent interruptions of the system, due to the fact that we are actively engaged in its improvement, the reader is invited to try it by registering at http://ssnt.l2f.inesc-id.pt.

## 8. CONCLUSIONS AND FUTURE WORK

This paper presented our prototype system for selective dissemination of broadcast news, emphasizing the impact of earlier errors of our pipeline system in the last modules. This impact is in our opinion an essential diagnostic tool for its overall improvement.

Our APP module has a good performance, while maintaining a very low latency for stream-based operation. The impact of its errors on the ASR performance is small (0.5% absolute) when compared with hand-labeled audio segmentation. The greatest impact of APP errors is in terms of topic segmentation, given the heuristically based approach that is crucially dependent on anchor detection precision.

Our ASR module also has a good real-time performance, although the results for European Portuguese are not yet at the level of the ones for languages like English, where much larger amounts of training data are available. The 51 hours of BN training data for our language are not enough to have an appropriate number of training examples for each phonetic class. In order to avoid the time-consuming process of manually transcribing more data, we are currently working on an unsupervised selection process using confidence measures to choose the most accurately anotated speech portions and add them to the training set. Preliminary experiments using additionally 32 hours of unsupervised annotated training data resulted in a WER improvement from 23.5% to 22.7%. Our current work in terms of ASR is also focused on dynamic vocabulary adaptation, and processing spontaneous speech, namely in terms of dealing with disfluencies and sentence boundary detection.

The ASR errors seem to have very little impact on the performance of the two next modules, which may be partly justified by the type of errors (e.g., errors in function words and in inflected forms are not relevant for indexation purposes).

Topic segmentation still has several pitfalls which we plan to reduce for instance by exploring video cues. In terms of topic indexation, our efforts in building better topic models using a discriminative training technique based on the conditional maximum-likelihood criterion for the implemented naïve Bayes classifier [27] have not yet been successful. This may be due to the small amount of manually topic-annotated training data.

In parallel with this work, we are also currently working on unsupervised adaptation of topic detection models and improving speaker clustering by using speaker identification. This component uses models for predetermined speakers such as anchors. Anchors introduce the news and provide a synthetic summary for the story. Normally, this is done in studio conditions (clean background) and with the anchor reading the news. Anchor speech segments convey all the story cues and are invaluable for automatic topic indexation and summary generation algorithms. Besides anchors, there are normally some important reporters who usually do the main and large news reports. This means that a very large portion of the news show is spoken by very few (recurrent) speakers, for whom very accurate models can be made. Preliminary tests with anchor speaker models show a good improvement in DER (droped from 26.1% to 17.9%).

## ACKNOWLEDGMENTS

The second author was sponsored by an FCT scholarship (SFRH/BD/6125/2001). This work was partially funded by FCT projects POSI/PLP/47175/2002, POSC/PLP/58697/2004, and European program project VidiVideo FP6/IST/045547. The order of the first two authors was randomly selected.

## REFERENCES

[1] H. Meinedo and J. Neto, "A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 237–240, Lisbon, Portugal, September 2005.

[2] R. Amaral and I. Trancoso, "Improving the topic indexation and segmentation modules of a media watch system," in *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP '04)*, pp. 1609–1612, Jeju Island, Korea, October 2004.

[3] I. Trancoso, J. Neto, H. Meinedo, and R. Amaral, "Evaluation of an alert system for selective dissemination of broadcast news," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH-INTERSPEECH '03)*, pp. 1257–1260, Geneva, Switzerland, September 2003.

[4] NIST, "Fall 2004 rich transcription (rt-04f) evaluation plan," 2004.

[5] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proceedings of DARPA Speech Recognition Workshop*, pp. 97–99, Chantilly, Va, USA, February 1997.

[6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proceedings of DARPA Speech Recognition Workshop*, pp. 127–132, Lansdowne, Va, USA, February 1998.

[7] J. Žibert, F. Mihelič, J.-P. Martens, et al., "The COST278 broadcast news segmentation and speaker clustering evaluation—overview, methodology, systems, results," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 629–932, Lisbon, Portugal, September 2005.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[9] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 2441–2444, Lisbon, Portugal, September 2005.

[10] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a broadcast news speech recognition system for the European Portuguese language," in *Proceedings of the 6th International Workshop on Computational Processing of the Portuguese Language (PROPOR '03)*, pp. 9–17, Faro, Portugal, June 2003.

[11] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," in *Proceedings of Automatic Speech Recognition: Challenges for the New Millenium (ASR '00)*, pp. 97–106, Paris, France, September 2000.

[12] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1281–1291, 2006.

[13] D. Williams, *Knowing what you don't know: roles for confidence measures in automatic speech recognition*, Ph.D. thesis, University of Sheffield, Sheffield, UK, 1999.

[14] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[15] S. Matsoukas, R. Prasad, S. Laxminarayan, B. Xiang, L. Nguyen, and R. Schwartz, "The 2004 BBN $1 \times$ RT recognition systems for English broadcast news and conversational telephone speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1641–1644, Lisbon, Portugal, September 2005.

[16] L. Nguyen, B. Xiang, M. Afify, et al., "The BBN RT04 English broadcast news transcription system," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1673–1676, Lisbon, Portugal, September 2005.

[17] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1149–1152, Lisbon, Portugal, September 2005.

[18] J. L. Gauvain, L. Lamel, and M. Adda-Decker, "Developments in continuous speech dictation using the ARPA WSJ task," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 1, pp. 65–68, Detroit, Mich, USA, May 1995.

[19] E. Shriberg, "Spontaneous speech: how people really talk, and why engineers should care," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1781–1784, Lisbon, Portugal, September 2005.

[20] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: identifying speaker role in radio broadcasts," in *Proceedings of the 7th National Conference on Artificial Intelligence and the 12th Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI '00)*, pp. 679–684, Austin, Tex, USA, July 2000.

[21] A. Gelbukh, G. Sidorov, and A. Guzmán-Arenas, "Document indexing with a concept hierarchy," in *Proceedings of the 1st International Workshop on New Developments in Digital Libraries (NDDL '01)*, pp. 47–54, Setúbal, Portugal, July 2001.

[22] Y. Y. Lo and J. L. Gauvain, "The LIMSI topic tracking system for TDT 2002," in *Proceedings of DARPA Topic Detection and Tracking Workhsop*, Gaithersburg, Md, USA, November 2002.

[23] S. Werner, U. Iurgel, A. Kosmala, and G. Rigoll, "Tracking topics in broadcast news data," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '02)*, Lausanne, Switzerland, September 2002.

[24] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: a success story," in *Proceedings of the Recherche d'Informations Assiste par Ordinateur (RIAO '00)*, Paris, France, April 2000.

[25] J. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "A system for selective dissemination of multimedia information resulting from the ALERT project," in *Proceedings of ISCA Workshop on Multilingual Spoken Document Retrieval (MSDR '03)*, pp. 25–30, Hong Kong, April 2003.

[26] C. Martins, A. Texeira, and J. Neto, "Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system," in *Proceedings of IEEE/ACL Spoken Language Technology Workshop*, pp. 146–149, Aruba, The Netherlands, December 2006.

[27] C. Chelba, M. Mahajan, and A. Acero, "Speech utterance classification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 280–283, Hong Kong, April 2003.

**R. Amaral** received the graduation and the M.S. degree in electrical engineering from the Faculty of Science and Technology of the University of Coimbra (FCTUC), Coimbra, Portugal, in 1993 and 1997, respectively. Since 1999, he is a Professor in the Electrical Engineering Department of the Superior Technical School of Setúbal-Polytechnic Institute of Setúbal, Setúbal, Portugal. He has been a researcher at INESC since 1997 at the Speech Processing Group that became the Spoken Language Systems Lab (L2F) in 2001. His Ph.D. topic is Topic Segmentation and Indexation on Broadcast News. He has participated in several European and National projects.

**H. Meinedo** graduated and received an MSc degree in Electrical and Computer Engineering from Instituto Superior Técnico (IST), Lisbon, Portugal in 1996 and 2000, respectively. He is finishing a Ph.D. degree also in Electrical and Computer Engineering from IST having as topic audio preprocessing and automatic speech recognition for Broadcast News. He has been a researcher at INESC since 1996 at the Neural Network and Signal Processing Group, that became the Spoken Language Systems Lab (L2F) in 2001. He has participated in several European and National projects.

**D. Caseiro** graduated in informatics and computer engineering in 1994 from Instituto Superior Técnico (IST), Lisbon, Portugal. He received an M.S. degree in electrical and computer engineering in 1998, and a Ph.D. degree in computer science, in 2003, also from IST. He has been teaching at this university since 2000, first as a lecturer, then as an Assistant Professor since 2004 (on compilers, and analysis and synthesis of algorithms). He has been a Researcher at INESC since 1996 at the Speech Processing Group that became the Spoken Language Systems Lab ($L^2F$) in 2001. His first research topic was automatic language identification. His Ph.D. topic was finite-state methods in automatic speech recognition. He has participated in several European and national projects, and currently leads one national project on weighted finite-state transducers applied to spoken language processing. He is a Member of ISCA (the International Speech Communication Association), the ACM, and the IEEE Computer Society.

**I. Trancoso** received the Licenciado, Mestre, Doutor and Agregado degrees in electrical and computer engineering from Instituto Superior Ténico, Lisbon, Portugal, in 1979, 1984, 1987, and 2002, respectively. She is a Full Professor at this university, where she lectures since 1979, having coordinated the EEC course for 6 years. She is also a Senior Researcher at INESC ID Lisbon, having launched the Speech Processing Group, now restructured as Spoken Language Systems Lab. Her first research topic was medium-to-low bit rate speech coding, a topic where she worked for one year at AT&T Bell Laboratories, Murray Hill, NJ. Her current scope is much broader, encompassing many areas in speech recognition and synthesis. She was a Member of the ISCA (International Speech Communication Association) Board, the IEEE Speech Technical Committee, and PC-ICSLP. She was elected Editor in Chief of the IEEE Transactions on Speech and Audio Processing (2003-2005), Member-at-Large of the IEEE Signal Processing Society Board of Governors (2006-2008), and Vice-President of ISCA (2005-2009). She chaired the Organizing Committee of the Interspeech'2005 Conference that took place in September 2005 in Lisbon.

**J. Neto** received his Graduation, M.S., and Ph.D. degrees in electrotechnical and computers engineering from the Instituto Superior Técnico, Technical University of Lisbon, in 1987, 1991, and 1998, respectively. He has been teaching at this university since 1991 where he is currently an Assistant Professor, teaching signal processing courses. He is a Researcher at INESC-ID Lisbon since 1987, and was one of the cofounders of the Spoken Language Systems Lab ($L^2F$) in 2000. His Ph.D. thesis was on speaker adaptation in a context of hybrid artificial neural networks and hidden Markov models continuous speech recognition systems. He has been working on these systems for broadcast news speech recognition applied to the Portuguese language. Also is working on the development of embodied conversational agents for different tasks. He was a Member of the Organizing Committee of the INTERSPEECH' 2005 Conference that took place in 2005 in Lisbon. He has participated in several European and national projects. He is a Member of IEEE and ISCA.