CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content

Martine Garnier-Rizet[1], Gilles Adda[2], Frederik Cailliau[3], Jean-Luc Gauvain[2], Sylvie Guillemin-Lanne[4], Lori Lamel[2], Stephan Vanni[1], Claire Waast-Richard[5]

[1]Vecsys, 3, rue de la Terre de Feu, 91952 Courtaboeuf Cedex, France{mgarnier,svanni}@vecsys.fr,
[2]LIMSI-CNRS,91403 Orsay cedex, FRANCE, {gadda,gauvain,lamel}@limsi.fr,
[3]Sinequa, 12, rue d'Athènes, 75009 Paris, France, cailliau@sinequa.com,
[4]Temis, Tour Gamma B, 193-197 rue de Bercy, 75582 Paris Cedex, France, sylvie.guillemin-lanne@temis.com,
[5]EDF R&D, Groupe SOAD, Dept ICAME, av. du général de Gaulle, 92141 Clamart Cedex 01, claire.waast-richard@edf.fr

Abstract

Being the client's first interface, call centres worldwide contain a huge amount of information of all kind under the form of conversational speech. If accessible, this information can be used to detect eg. major events and organizational flaws, improve customer relations and marketing strategies. An efficient way to exploit the unstructured data of telephone calls is data-mining, but current techniques apply on text only. The CALLSURF project gathers a number of academic and industrial partners covering the complete platform, from automatic transcription to information retrieval and data mining. This paper concentrates on the speech recognition module as it discusses the collection, the manual transcription of the training corpus and the techniques used to build the language model. The NLP techniques used to pre-process the transcribed corpus for data mining are POS tagging, lemmatization, noun group and named entity recognition. Some of them have been especially adapted to the conversational speech characteristics. POS tagging and preliminary data mining results obtained on the manually transcribed corpus are briefly discussed.

1. Introduction

Because of the economic implications linked to customer management, call centers have shown an exceptional growth during the last decade. This evolution corresponds to changes which have occurred in customer behavior and customer relationships. The call center is now a strategic interface for companies. For example, EDF sells energy and associated services to a broad clientele according to market segments (private, professional, SMEs). In order to answer its customers needs, EDF employs around 8000 agents and processes 25 millions of calls per year for the private market only. On a day to day basis, EDF and companies in general are concerned with how to well accommodate customers over the phone: limiting the waiting time, efficiently directing them towards the right representative or service. In order to succeed, marketing services need to better understand the purposes of these calls.

This can be done by carrying out frequent surveys, but these are costly and not conducted often enough. It can be done by analyzing fixed or free-form call-reasons fields. But the way these fields are fed can be unreliable. Therefore, qualitative studies analyzing some dialogues recorded and transcribed by hand can be done. However, conversational data in its volume remains still largely unexploited for obvious reasons of manual transcription cost. One solution is based on the automatic processing of the calls. But, until recently, the state-of-art in automatic speech transcription did not allow to imagine applications. The conversational telephone speech recognition was central in many DARPA evaluation campaigns. The situation is different in France where only few research laboratories and even less industrials work on the subject. As a result, there is no speech analytics system on the French market whereas the number of speech analytics solutions for American English is growing rapidly in the United-States. The purpose of CALLSURF which is part of the INFOM@GIC project is to adapt automatic speech recognition and text mining technologies for French, already applied successfully to other fields such as audio and video indexing, to conversational speech.

We will briefly present INFOM@GIC, a project conducted by Thalès. We will then describe the use case CALLSURF and its partnership. A large part will be dedicated to the call center data collection and manual transcription. We will follow with the work done on building the acoustic and linguistic models with an emphasis on disfluency analysis and processing. The current outcomes of text mining technologies such as morpho-syntactic labeling, named entities extraction and clustering will be described with a comparative approach on manual transcripts and speech decoder outputs. We will conclude on the evaluation work that will be conducted on the third and last period of INFOM@GIC-CALLSURF and the goals we want to reach at the end of the project.

2. Infom@gic – a Cap Digital Competitiveness Cluster project

INFOM@GIC is the major project of CAP DIGITAL[1] Competitiveness cluster. With a consortium of 29 partners leaded by Thalès, INFOM@GIC aims to identify and develop the key technologies in the fields of information analysis and knowledge extraction. At the end of the three years, INFOM@GIC's ambition is to provide professionals and individuals with advanced tools allowing to navigate easily and intelligently in very large amounts of data of any kind: Texts, Images, Sounds and Numerical data.

The interoperability is based on using the IBM UIMA platform. Among the INFOM@GIC use cases we may mention:

- Semantic Search on the web (PERTIMM)

[1] `http://www.capdigital.com/`

- Advanced functionalities for image search (EADS)
- Complex queries for urban videos archives management (EADS)
- INA's digitised patrimony (INA)
- Risks Detection (XEROX)

3. The CallSurf Use Case

CALLSURF belongs to the sub-project "Information extraction" and is dedicated to speech data. Its goal is to automatically identify the underlying purposes of customer call. This will help tremendously to spot trends, improve the effectiveness of quality assurance programs, reduce fraud, etc, without spending vast amounts of money on people to listen to and annotate calls. To reach such a goal, several options, with different underlying technologies, can be followed. Some people follow a keyword-phrase based approach that identifies predefined key words or phrases within the conversations. But it assumes that what is interesting to catch is already known. Others follow a phoneme-based approach. This allows a rapid adjustment to new languages but offers limited phonetic search functionalities. Finally one can follow a full orthographic transcription-based approach that attempts to transcribe all words of the conversations. This kind of speech analytics approach offers, from far, the largest potential of applications: from Quality Monitoring to customer issues identifying and understanding. In CALLSURF, we aim to be the first to bring to the French and to the European market such a high-level speech analytics suite. To do so, several steps are required. If the conversations between the parties are recorded on the same channel, automatic speaker segmentation and tracking keeps track of the speaker's identity. Then the automatic speech transcription performs a speech-to-text conversion. And finally, the information extraction engine mines and extracts, from transcriptions, the required information to index conversations and cluster them into categories. An overall functional architecture is given by Figure 1. The purpose of the CALLSURF project is to adapt existing technologies (speech recognition and text mining) to call center conversational speech. For French, if speech recognition technology has been successfully developed in broadcast news audio and video indexing domain ((Galliano et al., 2005), AUDIOSURF[2]) it has never been developed on real life human-to-human call center speech. This project was the opportunity to begin work in this domain and to improve speech recognition of call center data, including acoustic and language modeling adaptation. For text mining, usual modules like part-of-speech tagging, named entity and concept extraction, clustering, need to be adapted to speech input.

4. Corpus collection and transcription

The data collection and transcription still remain a necessary step to build the acoustic and the linguistic models. This work is particularly time consuming and consequently it is costly. In order to reduce the cost and in the same time increase the amount of transcribed data, we decided to build two kinds of corpus: a small corpus with fine transcripts, and a large corpus with fast transcripts. The main objective of such a step is to validate the use of fast transcripts as long as the volume is big enough.

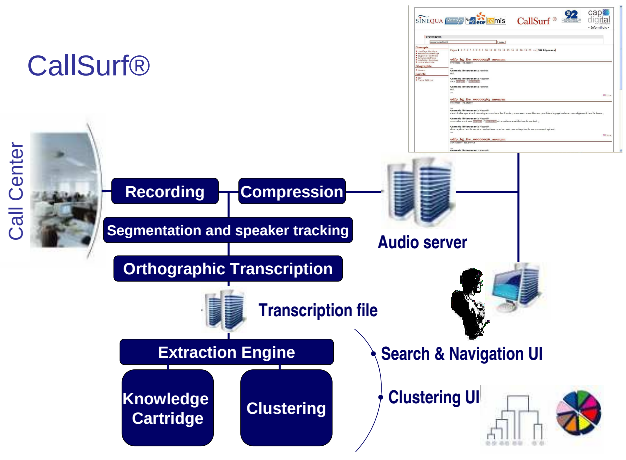


Figure 1: CallSurf functional scheme.

4.1. Audio Collection

During the 2006 summer, a Thales recording machine has been installed in one of the EDF Pro call centers. Four seats, which means a tenth of agents have been recorded, on a voluntary basis, during two months. It allowed us to collect around 620 hours of conversational data with a total of 5755 calls. As we explain further, we extracted 170 hours to build a 20 hours fine transcripts corpus (188 calls) and a 150 hours fast transcripts corpus (1268 calls).
Among the characteristics of this data, we could mention that:

- The agents were equipped with a headset microphone. The signal has been recorded at 64Kb/s (wav format).
- The overall quality is pretty good but it appears that parts of calls are noisy for different reasons (GSM, free-hands phone, noisy environment...)
- The duration of the calls goes from few seconds to more than half an hour with a mean of around 7 minutes. We eliminated the calls that were shorter than 15 seconds and longer than 30 minutes. The Figure 2 gives the call distribution according to the duration (for the 170 hours corpus).

We can notice that most of the calls last five minutes or less. The longest calls contain usually waiting music segments, silence segments corresponding to the customer folder access.

- The composition of the calls is heterogeneous: waiting music, recorded messages, telephone rings and of course, speech (dialogue and monologue),
- The Client and Agent have been recorded on the same channel. Consequently, overlapping speech appears.

  This aspect has caused many problems for the manual transcribers first and secondly for the automatic system. We will detail further how the overlapping speech

[2] http://audiosurf.limsi.fr/

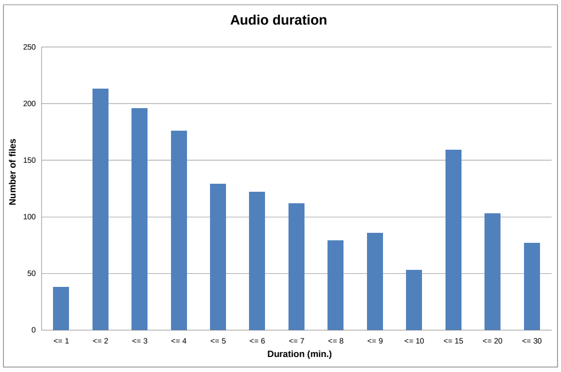


Figure 2: Duration distribution (1456 calls).

has been transcribed and processed. In the future, we think that the recording systems will take this feature into account for speech analytics purposes and record the two channels separately. Moreover, call centers will be more and more equipped with VoIP infrastructure which allows an easy separation of the two channels. It means that overlapping speech should not be a problem anymore in the near future.

- The collection is made of three kinds of calls (Agent→Client, Client→Agent, Client→Agent 1→Agent 2).

  The last type is made of calls in which the first agent contacts another one (a technical agent for example) and asks his customer to hold the line before coming back to him.

  Among the 1456 calls of the 170 hours corpus, the Client→Agent(s) calls represents more than 80% of which 30% are Client→Agent 1→Agent 2.

The Figure 3 shows the distribution of the different parts of calls (for the 20 hours corpus). The speech part represents 75% of the total duration with around 4% of overlapping speech. The ratio Agent/Customer is 60% / 40%.

4.2. Transcription procedures

The conversations we have been working on gave us a lot of precious information. In order to define the transcription conventions, we needed to get an overview of the content of such a real conversational speech corpus. We have added some punctuations (.,?) and case sensitiveness in all the transcribed corpora to approach a written structure, so as to facilitate the text mining processes.
From a transcription point of view we could describe the corpus as follows:

- On one side, the Customer speech is characterized by hesitations, disfluencies, and the use of non-professional terms and on the other side the Agent speech is less spontaneous, controlled, starting always with the same words `EDF Pro bonjour`.

- The vocabulary used by the agents between them is very specific with a lot of acronyms, procedure codes.

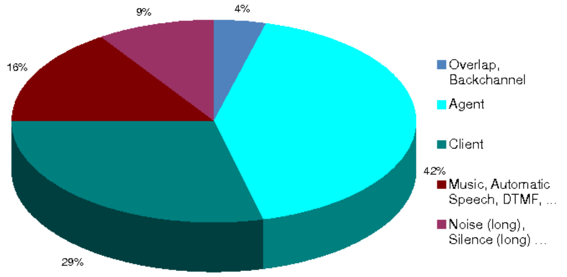


Figure 3: Distribution of different parts of calls.

- The calls contain a lot of proper names, addresses, EDF product names and it has been necessary to build a specific lexicon for the transcribers.

4.2.1. Anonymization procedure

The EDF Pro call center conversations contain personal data such as client name, company name, bank account, credit card number etc. that may not appear in the transcripts. Moreover, these data have no interest for the project since we focus on client concerns at large and not on individuals. That is why it has been decided to proceed to the anonymization of the transcripts.

4.2.2. Process of overlapping speech

As we have recorded on a single channel, we observe portions of overlapping speech. Two cases were distinguished:

- back channel overlappings (`hmm, ok, sure? ...`) are noted with a single and instantaneous tag, applied to the overlapped word (+`[bc]`)

- other portions are delimited by a turn and are not transcribed (the system is unable at this time to be trained from those portions)

4.2.3. Training corpora

As explained above, we built two sets of corpus according to our transcription goals:

- 20 hours fine transcription corpus: this corpus has been designed for training purpose, to construct acoustic models particularly. After defining specific conventions, we transcribed these 188 calls using Transcriber tool (Barras et al., 2001). For this data, specific tags were used for personal entities in order to facilitate the further anonymization procedure.

- 150 hours fast transcription corpus: the purpose of "fast transcripts" is to reduce the cost of manual transcription with the aim to get a larger amount of data with the same amount of money. As explained further this corpus has been largely used for building the acoustic models and for text mining purposes.

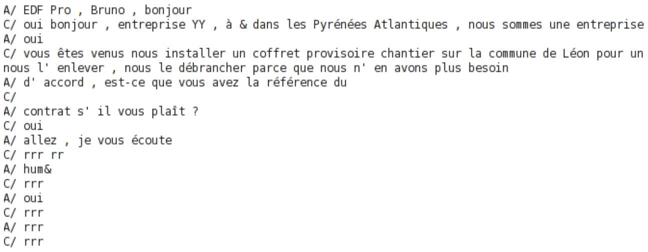
The main characteristics are:

- there is no text and signal alignment,

- we use a simple text editor,

- the personal items are directly anonymized (Mrs XX from the YY company, my bank account is cc c ccc cc ccc).

An example is given below and the Table 1 summarizes the differences between fine and fast transcripts.



```
A/ EDF Pro , Bruno , bonjour
C/ oui bonjour , entreprise YY , à & dans les Pyrénées Atlantiques , nous sommes une entreprise
A/ oui
C/ vous êtes venus nous installer un coffret provisoire chantier sur la commune de Léon pour un
nous l' enlever , nous le débrancher parce que nous n' en avons plus besoin
A/ d' accord , est-ce que vous avez la référence du
C/
A/ contrat s' il vous plaît ?
C/ oui
A/ allez , je vous écoute
C/ rrr rr
A/ hum&
C/ rrr
A/ oui
C/ rrr
A/ rrr
C/ rrr
```

Fine Transcripts	Fast Transcripts
• Text and signal alignment	
•Transcriber tool	•Simple text editor
•A posteriori anonymization	•On line anonymization
•Segment overlap. speech	•Transcribe overlap. speech
•Tagging of noise and breaths	
•Pronunciation information	
•Punctuation and case sensitiveness	•Punctuation and case sensitiveness

Table 1: Characteristics of fine and fast transcripts

4.2.4. Evaluation corpus

We have started to build a 10 hours of fine transcripts which will be used for a first evaluation of our work at different levels:

- to evaluate the first version of the speech recognition module adapted to EDF call center speech.
- to compare the tagging and clustering results on manual and automatic transcripts

5. Exploitation of the conversational corpus

5.1. Rich Automatic Transcription of call center conversational speech

The speech recognizer uses the same basic modeling and decoding strategy as in the LIMSI English broadcast news (Gauvain et al., 2002) and conversational telephone speech (Gauvain et al., 2005) systems.

5.1.1. Segmentation

Prior to transcription, audio segmentation is carried via an iterative maximum likelihood segmentation/clustering procedure (Gauvain et al., 1998; Barras et al., 2006). The procedure uses Gaussian mixture models (GMMs) representing speech and non-speech (i.e., silence or other the background conditions.) The result is a sequence of non-overlapping, acoustically homogeneous segments corresponding to the speaker turns in the audio document. The number of speakers is generally restricted to 2 for conversational speech.

5.1.2. Acoustic models

Each context dependent phone model is a tied-state, left-to-right context-dependent hidden Markov model (CD-HMM) with Gaussian mixture observation densities. The acoustic feature vector has 39-components comprised of 12 cepstrum coefficients and the log energy (estimated on a 0-3.8kHz band), along with the first and second order derivatives. Two sets of gender dependent, position dependent triphones are estimated using MAP adaptation (Gauvain et al., 1994) of SI seed models. The triphone-based context dependent phone models are word-independent but word position dependent. The first decoding pass uses a small set of acoustic models with about 5000 contexts and tied states. Larger sets of gender-dependent acoustic models covering 18k phone contexts represented with a total of 11.5k states are used in the latter decoding passes. State-tying is carried out via divisive decision tree clustering, constructing one tree for each state position of each phone so as to maximize the likelihood of the training data using single Gaussian state models, penalized by the number of tied-states (Gauvain et al., 1998). There are about 150 questions concerning the phone position, the distinctive features (and identities) of the phone and the neighboring phones.

5.1.3. Spectral normalization

VTLN is a simple speaker normalization at the front-end level which performs a frequency warping to account for differences in vocal tract length, where the appropriate warping factor is typically chosen from a set of candidate values by maximizing the test data likelihood based on a first decoding pass transcription and some acoustic models. Following (Hain et al., 1999), the MEL power spectrum is computed with a VTLN warped filter bank using a piecewise linear scaling function. The VTLN warping factors are estimated for each conversation side by aligning the audio segments with their word level transcription for a range of warping factors (between 0.8 and 1.25), using single-Gaussian gender dependent models (Welling et al., 1998) to determine the ML warping factor.

5.1.4. Recognition vocabulary and language model

A first word list for the speech recognition system was developed. This word list was obtained by selecting words from various text and transcriptions sources, in order to minimize the out of vocabulary (OOV) rate on a development set containing 63,000 words extracted from the CALLSURF fine transcripts. The OOV rate of this development data with a 40k word list is 0.8%. A 4-gram language model has been developed using the CALLSURF fast transcripts (1.4M words) and fine transcripts (0.2M words), together with texts and transcripts from the newspaper domain (600M words), and transcriptions of conversational telephone speech (3M words). The perplexity of the resulting 4-gram on the development set was 35, which is quite low but hides the intrinsic problems of the application: overlapping speech, conversational speech over the telephone channels, and anonymization.

5.1.5. Decoding

Decoding is carried out in two passes. In the first pass the speaker gender is determined for each conversation side, using Gaussian mixture models. Then a fast trigram decode is carried out generating an initial word hypothesis. Gaussian short lists and tight pruning thresholds are used to limit the computation time of the first decoding pass. This first hypothesis is also used to estimate the VTLN warp factors for each conversation side, and is used for acoustic model adaptation in the second pass using both the CMLLR and MLLR adaptation methods (Leggetter et al., 1995). MLLR adaptation relies on a tree organization of the tied states to create the regression classes as a function of the available data. This tree is built using a full covariance model set with one Gaussian per state. A word lattice is produced for each speech segment using a dynamic network decoder with a 2-gram or a 3-gram language model. This word lattice is rescored with a 4-gram language model and converted into a confusion network (using the pronunciation probabilities) by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in (Mangu et al., 1999), but appears to be significantly faster for large lattices. The words with the highest posterior in each confusion set are hypothesized along with their posterior probabilities.

5.1.6. Preliminary results and perspectives

A baseline Word Error Rate (WER) of about 47% was obtained using the LIMSI conversational telephone speech (CTS) transcription system. A first adaptation to the task was carried out by adapting the CTS acoustic models using the fine CALLSURF transcripts, and using all the available (fast and fine) transcripts for language modeling. The word error rate was reduced to about 35% on the development set, providing a first CALLSURF result. A large variability in word error rate was observed for the different conversation styles, and different speaker roles (higher WER on clients than on agents). The next task adaptation step will be the use the 150h of fast CALLSURF transcripts to adapt the acoustic models. This presents new challenges in that these transcripts do not contain any time information, that only a part of what was said is transcribed, and that 6% of the segments contains an anonymized item, but if we are able to overcome these problems, this technique may lead to a significant WER reduction.

In order to facilitate the work of the text mining modules, some rich transcriptions will be provided by developing methods to compensate the use of a automatic speech transcripts (add punctuation and case sensitiveness), the use of conversational speech (edit the disfluencies, structure the conversation), and to reduce the consequences of the intrinsic problems of the CALLSURF application (detection of OOV's and overlapping speech in the speech transcripts).

5.2. Text mining on conversational data: manual transcripts and decoder outputs

The process of text mining has a structure similar to the process of data mining with however additional requirements imposed by the specific properties of textual data expressed on natural language. There is a need to pre-process each textual document in order to reconstruct the missing data structure. Traditionally, this structure has a form of large dimension feature vector.

5.2.1. POS-tagging for conversational speech

One of the important pre-processing steps is the part-of-speech (POS) tagging. Each document is segmented in words and sentences. Segmentation is based on a mixture of language dependent rules and language lexicons. Each word is given a POS-tag using contextual rules for disambiguation. These were derived from a training corpus using a supervised training module (Brill, 1995). After this disambiguation, a lemma is provided by the lexicon. The tagger can load automatons to perform entity and concept extraction based on complex patterns. These may be adapted in a later phase of the project, as to limit the < performance drops that have been observed in the Audiosurf project (Cailliau et al., 2007).

For CALLSURF a specific language model for the part-of-speech tagger has been adapted to the call center speech data as to take into account speech specific lexicon and syntax. The 20h fine transcription corpus has been chosen as training corpus for this adaptation. Firstly, the training corpus has been tagged with a language model trained on written text (mainly news papers). All tags have then been manually verified and corrected. Some additional lexicons with energy related enterprises, abbreviations and speech specific vocabulary have been added to the tagger's knowledge base to limit the OOV occurrences.

Only POS-tags have been taken into account by the training module to create the disambiguation rules, limiting the number of learning features. Characteristics of the tagger are that it provides a POS-tag for every word, and that it does not disambiguate unless a disambiguation rule exists (multiple tags for a word are therefore possible). The performance of the new language model has been evaluated in terms of precision and recall, using the following measures:

Precision = total of correct tags/total of tags
Recall = total of correct tags/total of tokens

For the evaluation, the 98 files of the fine transcription corpus were arbitrarily distributed into 12 sets: 11 sets of 8 files and 1 set of 10 files. 12 language models have been created using 11 sets as training corpus and have been evaluated on the remaining set. Standard deviation, minimum, maximum and mean values for precision and recall are given in the following table:

	Min	Max	Mean	Standard deviation
Precision	0.9760	0.9853	0.9817	0.0080
Recall	0.9779	0.9860	0.9826	0.0074

These good results can probably be justified by a limited vocabulary and variety of syntactic structures, producing a language model that is less complicated than one trained on a written corpus.

The progression made by building a specific language model has been measured by tagging the fine transcription corpus with a language model trained on a journalistic corpus. This gave a mean precision of 0.9165 and a recall of 0.9986.

5.2.2. Text mining

To find out the underlying reasons for customer calls, we used a k-means approach applied on a feature vector extracted from the full conversation (20+150 hours transcription corpora). Each class is then characterized by a list of keywords that helps to identify the theme of its associated documents.

This was done through an automated classification server Insight Discoverer$^{\text{TM}}$Clusterer (so called IDC), which dynamically groups the feature vectors of documents. For a given document collection the IDC proposes the most relevant classification. It allows users to browse through their documents organized according to theme and sub-theme. A lexical filtering removing non relevant words and adding energy related concepts (specific terms, their acronyms and synonymous) was done preliminary. We used an iterative clustering with 10 clusters maximum by level and got 10 top clusters. Each of them were then subdivided into 4 to 10 sub clusters. The themes expressed in clusters descriptors reflect the customer main concerns: payment issues, contract cancellation due to a change in activity or moving, contract subscription...

Therefore, considering documents contents, their affectation to a given cluster remains unpredictable. Given a document, one can't decide which cluster will be automatically assigned to. This is due to two major reasons:

- Telephone calls contain a lot of repetitions and rephrasing, which may affect the weight of a theme.
- A lot of references are mentioned during the calls. The customer may have forgotten its contract number, or may spend some time to give a new address or a bank card number, which makes the duration of calls a lot longer, with a lot of mentions about references.

Concerning themes expressed in calls, it would be interesting to organize them in a hierarchy. Indeed the customer call concern has not the same status than the information which are exchanged during the call: contract reference, address, bank card number, etc.

Another approach would be to split calls in several sections depending on their topics, Different topic detection approaches, such as text-tiling, will be tested in the next phase of the project. A dedicated test corpus is under construction.

6. Conclusion

The first part of the INFOM@GIC-CALLSURF project allowed us: firstly to collect and transcribe "real" corpora from EDF call center with a total of 170 hours; secondly to adapt usual text mining techniques on conversational speech; and thirdly to consider two types of documents, the full conversation and customer turns as input for the clustering.

The second part of the project will be dedicated to two main goals:

- The development of rich transcription methods (punctuation and case sensitiveness, disfluencies edition, conversation structuring...) in order to make the text mining process easier.
- The implementation of evaluation procedures with the aim to
  - Measure the impact of recognition errors on the text mining modules (part-of-speech tagger, named entities extraction, clustering and classification), from a typology of the errors.
  - Measure the robustness of the language models (using data that are, at least, one year posterior than the training data, using data from another market sector).

7. Acknowledgments

This work was partly financed by the CAP DIGITAL Competitiveness Cluster project INFOM@GIC.

8. References

Barras, C., Geoffrois, E., Wu, Z., Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production *Speech Communication* **33**(1-2):5-22.

Barras, C., Zhu, X., Meignier, S., Gauvain, J.-L. (2006). Multi-stage speaker diarization of broadcast news, *IEEE Trans. Audio, Speech & Language Processing* **14**(5):1505-1512.

Brill. E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging *Computational Linguistics*, vol. 21, No. 4, 1-37.

Cailliau, F.& de Loupy, C. (2007). Aides à la navigation dans un corpus de transcriptions d'oral. TALN 2007. Toulouse. pp. 143-152.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G. (2005). The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News, Interspeech 2005.

Gauvain, J.-L., Adda, G., Lamel, L., Lefevre, F., Schwenk, H. (2005). Transcription de la parole conversationnelle, TAL, 45(3).

Gauvain, J.-L., Lamel, L., and Adda, G. (1998). Partitioning and transcription of broadcast news data. International Conference on Speech and Language Processing, vol. 4, pages 1335-1338, Sydney, Australia.

Gauvain, J.-L., Lamel, L., Adda, G. (2002). The LIMSI Broadcast News Transcription System, *Speech Communication*, **37**(1-2):89-108.

J.L. Gauvain, C.H. Lee, (1994). Maximum A Posteriori for Multivariate Gaussian Mixture Observation of Markov Chains, *IEEE Trans. Speech & Audio Proc.*, 291-298.

Hain, T., Woodland, P.C., Niesler, T.R., Whittaker, E.W.D. (1994). The 1998 HTK System for Transcription of Conversational Telephone Speech, *IEEE ICASSP,* Phoenix.

Leggetter, C.J., Woodland, P.C (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech & Language*, **9**(2):171-185.

Mangu, L., Brill, E., Stolcke, A. (1999). Finding Consensus Among Words: Lattice-Based Word Error Minimization, *Eurospeech'99*, 495-498 Budapest.

Welling, L., Haeb-Umbach, R., Aubert, X., Haberland, N. (1998). A study on speaker normalisation using vocal tract normalization and speaker adaptive training, *IEEE ICASSP*..