Classroom Lecture Recognition

Isabel Trancoso, Ricardo Nunes, and Luís Neves

INESC ID / IST,
R. Alves Redol, 9,
1000-029 Lisbon, Portugal,
Isabel.Trancoso@inesc-id.pt,
WWW home page: http://www.l2f.inesc-id.pt/~imt/

Abstract. The main goal of this work is to provide automatic transcriptions of classroom lectures for e-learning and e-inclusion applications. The first experiments using a recognition system trained for Broadcast News resulted in word error rates near 60%, clearly confirming the need for adaptation to the specific topic of the lectures, on one hand, and for better strategies for handling spontaneous speech. This paper describes the different domain adaptation steps that lowered the error rate to 45%, with very little transcribed adaptation material. It also includes a qualitative analysis of the different types of error, focusing on the ones related to a very high rate of disfluencies.

1 Introduction

The goal of the national project LECTRA is the production of multimedia lecture contents for e-learning applications. Nowadays, the availability on the web of text materials from University courses is an increasingly more frequent situation, namely in technical courses. Video recording of classes for distance learning is also a more and more frequent possibility. Our contribution to these contents (text books, slides, exercises, videos, etc.) will be to add, for each recorded video, the synchronized lecture transcription. We believe that this synchronized transcription may be specially important for hearing-impaired students.

This project encompasses 5 tasks. The first one concerns the collection of the training and test material (both in terms of recorded audio-video signals and textual data) related to a selected course. In the second task, this training data is used to adapt the acoustic, lexical and language models of our large vocabulary continuous speech recognizer (optimized for broadcast news transcription) to the course domain, thus yielding the automatic transcription of the lecture contents.

From a research point of view, the lecture transcription domain is very challenging, mainly due to the fact that we are dealing with spontaneous speech (mostly from the same speaker). Spontaneous speech is characterized by strong coarticulation effects, non-grammatical constructions, hesitations, repetitions, filled pauses, etc. [1]. For e-learning purposes, a plain transcription may not be intelligible enough, and may need ”enrichment” with punctuation, capitalization, marking of disfluencies, etc. These research challenges are the focus of the third task of the project.

The two last tasks of this project deal with integration and user evaluation, respectively. The goal is to integrate the recorded audio-video signal and the corresponding transcription with the other multimedia contents and synchronize them according to topic, so that a student may browse through the contents, seeing a viewgraph, the corresponding part in the text book, and the audio-video signal with the corresponding lecture transcription as caption.

For the user evaluation we intend to use both normal hearing and hearing impaired students, evaluating the lecture transcription with and without manual correction. This latter evaluation will give us an indication of how close we are in terms of automatic lecture transcription to be able to use such tools in real-time in a classroom.

Lecture transcription has been the target of much bigger research projects such as the Japanese project described in [2], the European project CHIL (Computers In The Human Communication Loop) [3], and the American iCampus Spoken Lecture Processing project [4]. In some of these projects, the concept of lecture is different. Our classroom lectures are almost 90 minutes long, and they involve mostly a single speaker (the teacher) who tried to create a very informal atmosphere. This contrasts with the 20 minute seminars used in [3], where a more prepared speech can often be found. Unfortunately, the amount of material for adapting our recognizer to the lecture domain is also very different from the very large amounts collected in other projects.

Section 2 summarizes the first task of the project - corpus collection, which started with two very different courses. Section 3 describes our baseline recognizer and the corresponding results. Section 4 is dedicated to the adaptation of the recognizer modules to the domain of the 2 courses. Section 5 summarizes our preliminary efforts in the third task in terms of dealing with the different sources of error we have encountered. Our last Section will discuss future research plans.

2 Corpora Collection

Two very different courses have been selected for our pilot study: one entitled "Economic Theory I" (ETI) and another one entitled "Production of Multimedia Contents" (PMC). Both were taught during one semester. The ETI course and the first 7 classes of the PMC course were recorded with a lapel microphone. The last part of the PMC course was recorded with a head-mounted microphone.

The two recording types presented specific problems. The lapel microphone proved inadequate for this type of recordings given the very high frequency of head turning of the teacher (towards the screen or the white board) that caused very audible intensity fluctuations. The use of the head-mounted microphone clearly improved the audio quality. However, the wireless communication system between the microphone and the sound recorder involved an automatic gain control, which actively increased the gain during the students questions, due to their distance from the microphone. As a result, the following reply from the teacher was highly saturated. Overall, 11% of the recorded segments with the head-mounted microphone were saturated.

No attempt was made to record the participation of the students in the class. The teachers were motivated to repeat their questions before answering them, but this was not frequently done.

The audio signal was extracted from the wmv (Windows Media Video 9 Codec) recordings using the ”VideotoAudioConverter” software, and converted to wav format at 16 kHz sampling rate, 16 bits per sample.

The recordings had variable duration, ranging from 40 to 90 minutes. All the ETI classes were taught by the same teacher, a male speaker with a Lisbon accent. Three of the PMC classes were taught by invited experts and the remaining 17 by the teacher, also a male speaker with Lisbon accent.

Due to very limited human resources, the manual transcription of all the classes of the two pilot corpora was totally infeasible. From the ETI course, we selected 3 segments of different classes, recorded several weeks apart, that served as training, development, and test sets. From the PMC course, we selected 3 other segments of distinct classes recorded by the main teacher, using the head-mounted microphone. The Transcriber software[1] was used for manually transcribing these segments. Table 1 shows the duration of the different sets for the two corpora and the number of words in each.

Table 1. Duration and number of words in each manually transcribed set.

	ETI			*PMC*		
	Train.	*Dev.*	*Test*	*Train.*	*Dev.*	*Test*
Duration (minutes)	62	46	36	73	52	42
Number of words	8k	7k	5k	10k	8k	6k

The very informal atmosphere of the PMC classes leading to highly spontaneous speech (even including laughter now and then) was not the only research challenge of this particular course. Because of its contents, it involved much computer jargon, usually derived from English (e.g. *email*, *software*), and a heavy use of spelt or partially spelt acronyms (e.g. *http*, *jpeg*). The computer jargon was generally pronounced very close to their English pronunciation, even introducing xenophones that are not part of the phone inventory for European Portuguese [5]. The percentage of technical terms in English in the PMC test corpus was 2.1%, a fact that will affect the lexical model, as we shall see below.

The ETI course also included some technical terms in English (e.g. *consumer price index*), but much more infrequently. The reference to mathematical variables and expressions, on the other hand, was much more frequent (e.g. $P1'$).

In order to adapt the language models to the domain of each course, we tried to get additional course materials (textbooks, viewgraphs, student reports, exam questions, etc.) which were first converted to text format and later processed by our text normalizer to expand abbreviations (KB - kilobyte(s), MHz - megahertz), numerals, etc. Finally, sentence boundary tags were added (<s>and </s >).

For the ETI course, we had a textbook and viewgraphs. Given the extension of the text book (452k words), we initially discarded the viewgraphs. For the PMC course, the textbook was in English, a frequent scenario in undergraduate engineering courses in

[1] http://trans.sourceforge.net

Portugal. So, in order to train language models in Portuguese we only had viewgraphs, exam questions and student reports. The text included in the viewgraphs amounted to 25k words, the exam questions to 2k words, and the student reports to 23k words.

Viewgraphs are typically characterized by specific grammatical constructions which clearly differentiates this material from other textual sources. By analyzing a small set of sentences from the PMC viewgraphs (around 2k words), the percentage of verbs that was obtained (9.1%) was much smaller then the one observed in a similar set of sentences from PMC reports (17.0%). The percentage of nouns, on the other hand, was much higher (42.2% in viewgraphs vs. 27.1% in reports). This different construction will have an obvious negative impact on the domain adaptation.

3 Baseline Recognizer

Our baseline large vocabulary recognizer was trained for Broadcast News (BN) in European Portuguese [6]. It uses hybrid acoustic models that try to combine the temporal modeling capabilities of hidden Markov models with the pattern classification capabilities of MLPs (Multi-Layer Perceptrons). The models have a topology where context-independent phone posterior probabilities are estimated by three MLPs given the acoustic parameters at each frame. The streams of probabilities are then combined using an appropriate algorithm [7]. The MLPs were trained with different feature extraction methods: 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 7 frames. The resulting network has a non-linear hidden layer with over 1000 units and 40 softmax output units (38 phones plus silence and breath noises). The language model was created by interpolating a newspaper text language model built from over 400M words with a backoff trigram model using absolute discounting, based on the training set transcriptions of our BN database (45h). The perplexity (PP) is 139.5. The vocabulary includes around 57k words. For the BN development test set corpus, the out-of-vocabulary (OOV) word rate is 1.4%. The lexicon includes multiple pronunciations, totaling 66k entries.

For Broadcast News, this baseline recognizer achieves an average WER (word error rate) for all conditions of 32%, which decreases to 13%, in F0 conditions (read speech, studio recordings). These results were obtained at 4.4x real time in a Pentium IV Processor at 2.66GHz.

3.1 Audio Segmentation

The BN recognizer has a pre-processing module that performs audio segmentation [7]. This module segments the audio file into homogeneous regions according to background conditions, speaker gender and special speaker id (anchors). The speech/non-speech classification is used in the lecture transcription to exclude from the recognition task the segments containing questions or comments from the students. These segments are recorded with a very distant microphone, which makes them almost unintelligible in many cases. Besides excluding almost all these segments, the pre-processing module also excludes some very noisy segments with the teacher's voice.

We had about 170s with student contributions in each of the two test sets. The classification module excluded almost all of these segments (except for 9s in the ETI test set and 3s in the PMC test set).

3.2 Recognition Results

The BN recognizer described above was first applied to the transcribable segments, without any adaptation. The WER was 56.4% and 63.6%, respectively, for the ETI and PMC test sets. These very bad results were expected in view of the fact that we are dealing with spontaneous speech recorded in a classroom, with very specific contents. Furthermore we had to cope with recording problems related either to head shifts relative to microphone positioning in one case, or to very frequent saturation effects in another. The lack of domain adaptation is specially patent in the high OOV rates and perplexity values obtained for the PMC test set (OOV=3.4%, PP=292.8). For the ETI test set, the values were much lower (OOV=1.6%, PP=175.0).

4 Domain Adaptation

The following subsections describe the adaptation stages to the lecture domain of the lexical, language and acoustic models. We tried to make this process as automatic as possible in order to enable the rapid porting to other course domains.

4.1 Lexical Model

For one of the two courses selected for our pilot study (PMC), the simplest approach of adding new entries to the pronunciation lexicon by running them through an automatic grapheme-to-phone conversion module for European Portuguese [8] would not be advisable, given the high percentage of technical terms of English origin.

Our first attempt consisted of designing a set of hand-crafted rules to separate the words that would likely be of foreign origin. The regular expressions dealing with grapheme sequences were written using the flex program and achieved a correct identification rate of 65% on the PMC viewgraphs training set. Given the high miss rate, we tried the intersection with an English lexicon of approximately 118k entries, and a total of 127k multiple pronunciations.[2]

This procedure was followed by some phone mapping, from the English phone inventory to the European Portuguese one (using SAMPA). At this stage we have excluded the use of xenophones. In the context of lectures for undergraduate university students, foreign technical terms are most often pronounced fairly close to their English pronunciation (e.g. "position" would be pronounced as [pɐziʃɐn] or [poziʃɐn] instead of the Portuguese pronunciation [pozitjõ]). The variability caused by the possible degrees of nativization is still enormous [9]. The phone mapping between the English phone inventory and the Portuguese one may not be unique. As an illustration, take the English phone [θ], a xenophone which can be pronounced either as [s] (closest symbol in terms

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

of pronunciation) or [t] (closest symbol in terms of orthography, since *th* sequences never occur in Portuguese).

The new PMC vocabulary includes around 3k new entries, of which 80.4% correspond to technical terms in English or acronyms. Like foreign words, acronyms are a special class that does not follow the common lexicon rules. The rules for spelling acronyms are trivial, but partially spelt or read acronyms in European Portuguese have much more complex rules and are characterized by a high degree of pronunciation variability, even among native speakers. Nowadays, we can also find many acronyms in the mention to URLs.

For the ETI course, we have selected for the vocabulary all the words of the transcribed training material, plus the words of the text book that occurred more than 5 times. The new ETI vocabulary includes around 325 new entries, of which 17.5% correspond to acronyms. The most frequent OOVs of the ETI test set are references to mathematical variables (33.8%).

4.2 Language Model

Given the scarcity and inadequacy of written training material for the PMC pilot course, building a language model on that basis alone would give rise to a very high perplexity (256.1) and OOV rate (8.0%). The best results were hence obtained by interpolating the new language model with the one derived from the broadcast news domain. The new 3-gram language model was built using the SRILM toolkit [10], with modified Knesser-Ney discounting. Before interpolation, the WER corresponding to this new model was 64.8%. After interpolation, it decreased to 58.7%. The perplexity decreased to 208.6 and the OOV rate to 1.7%.

For the ETI course, the interpolation of the textbook and the transcribed training lecture with the BN model decreased the WER to 54.3%, corresponding to a perplexity of 127.7. The OOV rate was practically the same as with the initial BN model.

4.3 Acoustic Model

Although the transcribed training material was also very scarce, it was worth testing how much one could gain from adapting the acoustic models to the speaker and classroom environment, with just one lecture. The adaptation procedure was slightly modified to take into account that we had no initial models for filled pauses. We tried adapting the acoustic models without and with language model adaptation. In the first tests, after 3 iterations, the WER was down to 48.0%, for the PMC course and to 45.4% for the ETI course. The WER reduction was hence much more significant than with language model adaptation alone. In the second tests, the WER decreased to 44.8% for the PMC course, and to 44.7% for the ETI course.

5 Error Analysis

A clear course for improving the word error rate is to get more training data, namely in terms of additional transcribed material. Another course is to make a qualitative and quantitative error analysis, hoping that our small test set would be representative enough to indicate typical error sources. This section is our first step in this direction. The qualitative analysis of the errors for the two test sets indicates the following types of error:

- 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) intra-word 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 not specific of spontaneous speech, being strongly affected by factors such as high speech rate. The relatively high deletion rate may be partly attributed to severe vowel reduction. It is also worth noting that 61.4% of the deleted words in the PMC test set are (typically short) function words.
- 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 post-processing 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. It is known that one OOV term can lead to between 1.6 and 2 additional errors [11]. 32.0% of the OOVs in the PMC test set are verbal forms. The current lexicon does not have too many verbal forms with clitics (e.g. *desenvolveu-se*, 'developped'). It may be worth exploring the possibility of separating clitics when building the lexical and language models, although our previous attempts of doing some morphological analysis have not yet brought any significant improvements [12].
- Errors around speech disfluencies. This is the type of error that is most specific of the spontaneous speech of our lecture corpus. The frequency of repetitions, repairs, restarts and filled pauses is very high, in agreement with values of one disfluency every 20 words cited in [1]. Unfortunately, the training corpus for Broadcast News included a very small representation of such examples, and our manually transcribed corpus was far too small.
- Errors in tag questions. This type of construction is fairly frequent in both courses, given the need felt by the teachers to make sure that the class was following their presentations. Therefore the teachers often invited the students to give feedback by using tag questions such as *não é?* ('isn't it?', 36 instances in the ETI test set), *(es)tá bem?* ('all right?'), *(es)tão a ver?* ('are you understanding?'), or the nativized version of *okay?*. The very casual articulation of these words, coupled with the virtual non-representiveness of such examples in the written corpus makes them very difficult to be recognized.

The type of errors around disfluencies is the one that mostly differentiates this corpus from other corpora we have worked on, and therefore deserves our particular attention. At this early stage, it was not yet possible to make a quantitative analysis of all type of fillers [13] found in our corpus: filled pauses, discourse markers, explicit editing terms and asides/parentheticals. So far, we have mostly concentrated on filled pauses which are the easiest ones to automatically detect.

Filled pauses are hesitation sounds that speakers employ to indicate uncertainty or to maintain control of a conversation while thinking of what to say next. They can occur anywhere in the stream of speech. For European Portuguese, following the proposal of [14], the most common filled pauses were transcribed as *mm* (when the sound is produced with a closed mouth, sounding either as [ɨ̃ː] or [ũː]), *aam* (when there is evidence either of a nasal vowel or a vowel followed by a nasal murmur, sounding as [ɨ̃ː], [ɐ̃] or [ɐ:m]), and *aa* (when the sound corresponds to a non-nasal vowel, similar to either [ɨː] or [ɐ:]). The most common filled pause in our corpus of lectures is by far *aa*, very similar to the Portuguese article and preposition *a*, one of the top 5 most frequent words.

In the PMC test corpus, 1.9% of all manually transcribed words were filled pauses. In the ETI test corpus, the percentage was much lower (0.4%), although it had higher values in the training and development sets.

For filled pause detection, we implemented a method based on the relative stationarity of F0 and spectral slope over filled pauses. This method, which was far simpler but not as efficient as the one described in [15], has a 35.2% recall rate (number of filled pauses detected correctly / total number of filled pauses) and a 99.9% precision rate (number of filled pauses detected correctly / total number of filled pauses detected). This was the first step towards building acoustic models for filled pauses, which were not previously included in our BN recognizer.

Repetitions are also relatively easy to detect, when the repeated material is exactly the same as originally pronounced (29 instances of one-word and two-word repetitions in the TEI test set and 22 for PMC), but more complex revisions are more frequent.

Discourse markers are also extremely frequent in our corpus. The most typical discourse marker is by far *portanto* ('so'), which was pronounced in many different ways, mostly in very reduced forms, in both courses (104 instances in the ETI test set and 42 in the PMC test set). In fact, the recognition error of this particular discourse marker was very high (71.4% in the ETI test set and 60.2% in the PMC test set). The nativized version of *okay* was also fairly frequent in the PMC course. In the ETI course, on the other hand, we could find many other examples of discourse markers, such as *ora bem* ('well', 7 instances, 57.1% error rate), and *reparem* ('notice', 26 instances, 69.2% error rate). This variability in using discourse markers is very speaker and dialect dependent.

The individual discourse style differences are also very interesting. For instance, the ETI teacher uses rethorical (or hypothetical) questions very often, whereas the PMC teacher prefers statements or questions to the audience.

A significant part of the recognition errors occurs for function words. In the ETI test set, 44.5% of all words are function words, and the percentage is similar for the PMC test set (45.0%). 47.3% of all recognition errors in the ETI test set occur for function words, and 42.9% in the PMC test set. These results make us believe that the current performance, although too bad in terms of transcription, may be good enough for indexation purposes. This is specially important for the lecture browsing application, but this feature has not yet been included.

Error bursts, i.e. sequences of wrongly recognized words, were fairly frequent (e.g. around disfluencies). In the PMC test set, only 19.2% of the errors occurred in isolation; 18.6% occurred in bursts of two errors; 53.2% in bursts of 3-9 errors; and 8.9% in longer bursts. Similar statistics could be found for the ETI test set.

In the above analysis, we have not mentioned errors due to inconsistent spelling of the manual transcriptions, which were, however, relatively frequent. The most common inconsistency consists of writing the same entries both as separate words and as a single word (e.g. colormap and color map).

6 Conclusions and Future Work

This pilot study with lecture transcription allowed us to learn valuable lessons in terms of recording protocols, and validated the well known importance of large quantities of textual and manually transcribed material for training language and acoustic models. Despite the very limited resources, our domain adaptation efforts yielded a significant (although not sufficient) word error rate reduction.

Further error reductions must be obtained at the cost of better strategies for dealing with disfluencies. However, some of the identified error sources are not exclusive to spontaneous speech recognition. In fact, we are currently dealing with them in the scope of our efforts for automatic captioning of broadcast news. We believe that the use of much larger speech and text corpora may obviously decrease these problems, namely by using context-dependent acoustic models, but much can be gained by studying phenomena such as vowel reduction.

Producing a rich transcription for lectures does not only entail dealing with disfluencies, but also punctuation and capitalization, which are the focus of another PhD thesis in the scope of this project. The research challenges are enormous, not only in terms of disfluency detection and repair [16] [17] [18], but also in terms of producing a surface rich transcription [19] that is more intelligible for hearing impaired students.

New corpora have already been recorded in the current semester, not only for e-learning purposes, but most specially for helping an undergraduate student with progressive hearing disabilities. We are currently dealing with the additional challenges posed by a course on Algebra, which is particularly interesting, as it involves mentioning mathematical variables and expressions.

It will be worth investigating whether some time savings in terms of speaker adaptation can be achieved by asking the teacher to record a (read) technical text prior to starting the course. The manual transcription of the lectures themselves, however, remain very important for researching spontaneous speech phenomena, in particular speaker-dependent disfluencies [18]. It will also be worth investigating whether the existence of transcribed data for one course can be beneficial for another one, in a totally different domain. This would mean building a corpus of University lectures for several courses.

Acknowledgment The authors would like to thank Ciro Martins, Hugo Meinedo, João Neto and Céu Viana for many helpful discussions. This work was partially funded by

FCT project POSC/PLP/58697/2004. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

References

1. Shriberg, E.: Spontaneous speech: How people really talk, and why engineers should care. In: Proc. Interspeech '2005, Lisbon, Portugal (2005)
2. Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., Tamura, S.: Ubiquitous speech processing. In: Proc. ICASSP '2001, Salt Lake City, USA (2001)
3. Lamel, L., Adda, G., Bilinski, E., Gauvain, J.L.: Transcribing lectures and seminars. In: Proc. Interspeech '2005, Lisbon, Portugal (2005)
4. Glass, J.R., Hazen, T.J., Hetherington, I.L., Wang, C.: Analysis and processing of lecture audio data: Preliminary investigations. In: Proc. Human Language Technology NAACL, Speech Indexing Workshop, Boston (2004)
5. Lindström, A.: English and Other Foreign Linguistic Elements in Spoken Swedish: Studies of Productive Processes and Their Modelling Using Finite-State Tools. PhD thesis, Linköping University (2004)
6. Trancoso, I., Neto, J., Meinedo, H., Amaral, R.: Evaluation of an alert system for selective dissemination of broadcast news. In: Proc. Eurospeech '2003, Geneva, Switzerland (2003)
7. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: Proc. ICASSP '2003, Hong Kong (2003)
8. Caseiro, D., Trancoso, I., Oliveira, L., Viana, C.: Grapheme-to-phone using finite state transducers. In: Proc. 2002 IEEE Workshop on Speech Synthesis, Santa Monica, CA, USA (2002)
9. Trancoso, I., Viana, C., Mascarenhas, M., Teixeira, C.: On deriving rules for nativised pronunciation in navigation queries. In: Proc. Eurospeech '1999, Budapest, Hungary (1999)
10. Stolcke, A.: Srlim - an extensible language modeling toolkit. In: Proc. ICSLP '2002, Denver, USA (2002)
11. Gauvain, J., Lamel, L., Adda, G.: Developments in continuous speech dictation using the arpa wsj task. In: Proc. ICASSP '1995, Detroit, USA (1995)
12. Martins, C., Neto, J., Almeida, L.: Using partial morphological analysis in language modeling estimation for large vocabulary portuguese speech recognition. In: Proc. Eurospeech '1999, Budapest, Hungary (1999)
13. LDC: Simple metadata annotation specification version 6.2. Technical report, Linguistic Data Consortium (2004)
14. Mata, A.: For a Study of Intonation in Spontaneous and Prepared Speech In European portuguese: Methodology, Results and Didactic Implications (in Portuguese). PhD thesis, FLUL, Lisbon (1998)
15. Goto, M., Itou, K., Hayamizu, S.: A real-time filled pause detection system for spontaneous speech recognition. In: Proc. Eurospeech '1999, Budapest, Hungary (1999)
16. Heeman, P., Allen, J.: Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. Computational Linguistics **4**(25) (1999) 527–571
17. Johnson, M., Charniak, E.: A tag-based noisy channel model of speech repairs. In: Proc. ACL, Barcelona, Spain (2004)
18. Honal, M., Schultz, T.: Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies. In: Proc. ICASSP '2005, Philadelphia, USA (2005)
19. Snover, M., Schwartz, R., Dorr, B., Makhoul, J.: Rt-s: Surface rich transcription scoring, methodology, and initial results. In: Proceedings of the Rich Transcription 2004 Workshop, Montreal, Canada (2004)