# Summarizing Spontaneous Speech Using General Text Properties

**Maria Fuentes, Edgar Gonzàlez,**
**Horacio Rodríguez, Jordi Turmo, Laura Alonso**

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
{mfuentes,egonzalez,horacio,turmo,lalonso}@lsi.upc.edu

## Abstract

In this paper we describe and analyze the performance of a speech summarization prototype based in classical text summarization techniques, mainly lexical chains. We show that shallow text-oriented techniques are robust enough to identify relevant information and basic units of content in spontaneous speech. Moreover, the performance in recovering the most relevant segments does not drop with noisy input, as in automatic transcripts.

Segments from transcripts are extracted taking into account information provided by lexical chains, shallow parsing, collocations and discourse markers. Two different approaches to segmentation are evaluated. In the PreSeg approach, as in classical textual summarizing, segmentation is part of the preprocessing step, while in the PostSeg approach, summary text segments are found after the detection of relevant document parts.

## 1 Introduction

With the increasing importance of human-machine interaction, speech, as the most natural way of human communication, has become the core of many computer applications. Moreover, automatic summarization is of help in digesting the increasing amounts of information that reach us every day. Thus, it seems clear that the combination of speech and summarization would improve the interaction between humans and computers. This work is part of the CHIL project[1], aimed to provide intelligent solutions for human information needs.

Leaving aside the problems intrinsic to summarization, the main problem in summarizing oral input is that the performance of automatic speech recognizers (ASRs) is still far from perfect, and so words are often recognized wrongly (e.g.: *gate* by *Kate*), thus misleading all subsequent linguistic processing.

Most of the work in summarization of oral speech has focused in broadcast news, which are typically read aloud from a text. The performance of ASRs is better in read text, and the linguistic register is closer to the typical input of automatic summarizers, written text.

However, summarization of spontaneous speech is also useful, for example, to grasp the main idea of political speeches, of conferences or to quickly review corporate meetings. In these cases, the problems of automatic recognition must be added to the problems caused by the fact that the linguistic register is very different from standard, written text: there is no punctuation or capitalization, utterances are often ungrammatical or ill-formed (they contain disfluencies, repairs or unfinished sentences), information tends to be more redundant than in written text, etc. Despite these difficulties, (Koumpis & Renals 00) summarize voice mail, and (Kikuchi *et al.* 03) deal with presentations.

To tackle the variation in register and the noise of faulty ASRs, robust approaches to summarization are needed. In this study, we exploit lexical chains and discourse markers to detect relevant information in transcripts of monologues. Acoustic information has not been considered because it is not always available. We believe that these techniques are specially suitable for the problem of summarizing spontaneous speech, because in oral presentations important concepts tend to be highly repeated and discourse markers are frequently used.

The success of (Stokes *et al.* 04) in exploiting lexical chains to obtain short summaries of

[1] chil.server.de/servlet/is/101/

broadcast news seems to indicate that this approach is robust enough to deal with different registers.

We have produced 10- and 30-word extractive, indicative summaries of transcripts of presentations at the EuroSpeech'93 conference, obtained from the Translingual English Database (TED) corpus[2]. Automatic summaries have been evaluated by comparison with the title and keywords of the paper and with extractive summaries produced by two human judges. The ROUGE (Lin 04) package has been used for evaluation, to assess similarities in content by unigram overlap. We have found that automatic summaries from a presentation are reasonably similar (content-wise) to human ones, and also to the title and the keywords of the corresponding paper. The evaluated summarizers perform significantly better than two dummy baselines. The first baseline is the first fragment of the talk, where the speaker usually synthesizes the aim of the talk. The second baseline consists in selecting the speech segment that maximizes the total frequency score, summing up the frequencies of all the words in it.

The rest of the paper is structured as follows. In the next section the architecture of the summarizers is detailed. In Section 3 we describe the method, the data set and the conditions used for evaluation. Section 4 discusses experiments and results, and we finish with some conclusions.

## 2 Architecture

We present here two summarization prototypes, **PreSeg** and **PostSeg**, whose architectures (see Figure 1) share a highly portable core. This core relies on domain and register independent linguistic processes, and specific modules are added as required.

In written text summarization, sentence-like segments (as opposed to topic-based segments, virtually not used for extractive summarization), are usually identified by punctuation marks, but this simple approach is unfeasible in spontaneous speech due to the lack of punctuation. **PreSeg** and **PostSeg** differ in the way they tackle the task of segmenting text: in the **PostSeg** approach the identification of segment boundaries is deferred until relevant content is detected, whereas in the **PreSeg** approach the text is segmented at the preprocessing step, as detailed below.

Two external resources are required for segmentation: a list of collocations and a list of discourse markers[3] (Alonso Alemany 05). Collocations[4] are extracted automatically from a written corpus applying a $\chi^2$ hypothesis test (as proposed in (Manning & Schütze 99)) to all n-grams of up to a certain size. In the **PostSeg** approach only collocations filtered by syntactic patterns are considered (a subset of those in (Arranz *et al.* 05), see table 1). Every discourse marker is associated to a score of rhetorical relevance (2, 1 or -1). This score is assigned to all words within the scope of the discourse marker (following or preceding the marker, depending on its syntactic type).

The main tasks performed by the prototypes are the following:

1. Preprocessors: They perform generic NLP preprocessing tasks over an input text (tokenizing, tagging, lemmatizing, syntactic chunking and semantic labeling, identification of discourse markers and collocations) [5].

In the **PreSeg** approach, the input text is segmented at this point, as detailed in the following item. In the case of **PostSeg**, the decision is deferred.

1a. PreSeg Transcript Segmenter: Segments of $n$ words are identified in the input text, with the restriction that segment boundaries cannot be placed

- before a coordinating conjunction,
- after a conjunction, a preposition or a determiner,
- so that they split a syntactic chunk or a collocation,

[2] `www.elda.org/catalogue/en/speech/S0120.html`

[3] `http://russell.famaf.unc.edu.ar/~laura/shallowdisc4summ/discmar/`

[4] sequences of words that are likely to co-ocur.

[5] In most cases, no adaption has been done and the preprocessors are based on written text models.

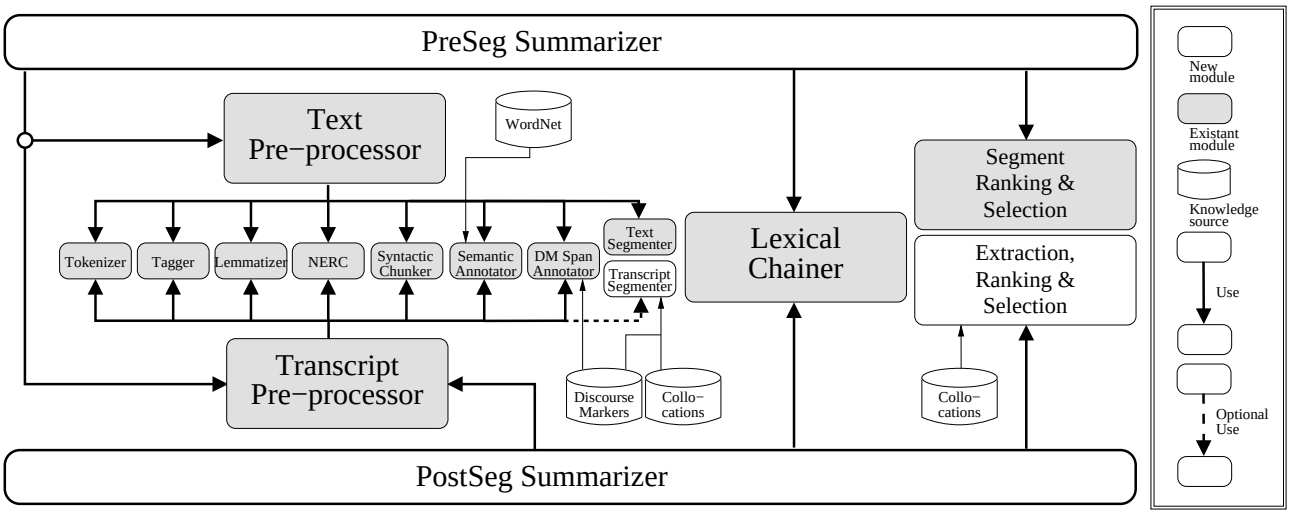


Figure 1: Summarizer Architecture

| |
|---|
| definite−{JJ}_clause−grammar{NN} |
| non-iterative{JJ}_matrix−inversion{NN} |
| bottle{NN}_necks{NNS} |
| gradation{NN}_of{IN}_spontaneity{NN} |
| bionic{JJ}_wizard{NN} |
| spaces{NNS}_of{IN}_perceptual{JJ}_distinction{NN} |
| correlates{VBZ}_of{IN}_perceptually{RB} |

Table 1: Sample of collocations filtered by syntactic patterns.

– so that they leave a discourse marker in the last $m$ positions of the segment (where $m$ is related to the scope size of the discourse markers).

When a discourse marker is found and we need less than $m$ words to complete a segment, a boundary is placed immediately before it, yielding a segment shorter than $n$ words. As for the rest of restrictions, while they cannot be satisfied more words are added to the segment, until a suitable splitting point is found, possibly producing segments longer than $n$.

2. Lexical Chainer: First, lexical chains are identified: a lexical chain is constituted by all the occurrences of semantically related lexical items in the text. WordNet (Fellbaum 98) is used to identify synonymy relations between words. In contrast to (Stokes *et al.* 04), our algorithm takes a non-greedy approach. The proper sense, or a reduction of the number of possible senses of ambiguous words, is determined when all words in the document have been processed. So, lexical chain identification performs a reductive Word Sense Desambiguation process as well. Once chains are identified, they are scored according to a number of heuristics: their length, the kind of relation between their words, their starting position within the text, etc. Chains are then classified into strong, medium and light, depending on their score:

$$\begin{aligned} \tau &= \mu_s + 2 \cdot \sigma_s \\ Strong &= \{c \mid score_c \geq \tau\} \\ Medium &= \{c \mid \tau > score_c \geq \tau/2\} \\ Light &= \{c \mid \tau/2 > score_c\} \end{aligned}$$

where $\mu_s$ and $\sigma_s$ are the average of the scores of all the chains and the corresponding standard deviation. In order to recover most of the information in the text, we consider medium and light chains, instead of only strong ones. In spontaneous speech, strong chains tend to provide a misrepresentation of the information

in a text, because the distribution of the frequency of words is rather squewed, and only few strong chains are found.

3. Summary Generation

3a. Segment Ranking and Selection (PreSeg approach): Segments crossed by the most highly scored lexical chains are ranked according to their order of occurrence, so that the sooner they occur in the text, the higher they are ranked. The smallest set of best ranked segments comprising a number of words larger or equal to the targeted size is taken as the summary.

3b. Extraction, Ranking and Selection (PostSeg approach): Taking into account the lexical chains found in the previous step, this module extracts windows of $n$ contiguous words (chunks) to form a summary of the targeted size.

Chunks are included in the summary using a priority function that tries to capture both relevance and well-formedness. First, the most top-priority chunk covering the first position where the most highly scored chain occurs is selected, because the first occurrence of a relevant word is believed to present the concept and relate it to other relevant concepts. In the following steps, the algorithm selects the following candidate chunk with highest priority whose content does not overlap with that of the already selected chunks, until the number of selected words is equal or greater than the targeted summary size.

Chunks are ranked by the following criteria (from most to least discriminating):

1. Least internal repetitions (occurrences of the same chain inside a chunk)
2. Highest sum of covered chain scores.
3. Highest number of strong lexical chains.
4. Highest number of new (not present in already selected chunks) chains.
5. Lowest collocation breakage (sum of the weight, statistically determined with $\chi^2$, of all the collocations the chunk boundaries break).
6. Lowest number of violations on the conditions of well-formedness of chunks: restrictions on the PoS of the starting and final words, breakage of syntactic chunks (as in item **1a.**).
7. Highest rhetorical relevance (sum of the rhetorical relevance of all words in the chunk, determined by the presence of discourse markers).
8. Highest size suitability (1 for chunks whose addition would give a summary with exactly the desired size, $-1$ for those whose addition would leave a gap smaller than the minimum chunk size).
9. Earliest Starting Position.
10. Earliest Ending Position.

These criteria try to determine the priority class chunks in a several step strategy: criteria 1 to 4 try to determine regions of relevance within the document, finding a set of chunks covering relevant text fragments. Criteria 5 and 6 are intended to refine the selection and to find, among the set of relevant chunks, those which are also syntactically better-formed. Criteria 7 tries to exploit text coherence. The last three criteria are aimed to break ties between any remaining candidates, taking into account chunk size and position.

When one chunk is added to the summary, the score of the lexical chains occurring in the chunk is reduced to its half. This causes the scores used in criterion 2 to readjust for the following calculation of top-priority segments, making it more robust against the redundancy of spontaneous speech.

## 3 Evaluation framework

Summary content evaluation is known to be a difficult task. Objective judgments are needed to assess the progress achieved by different automatic approaches, typically requiring a costly human effort. Some contests have been carried out to evaluate summarization systems with common, public procedures, such as Document Understanding Conference (DUC) (Over & Yen 03), where different sets of criteria have been provided to evaluate summary quality. The evaluation should show the system's performance both in content selection

and grammaticality. However, as we work with an ill-formed input (often ungrammatical), we have limited our evaluation to a quantitative analysis on content.

### 3.1 Automatic Evaluation method

To establish the quality of the summarizers we decided to use measures from ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package. For single document summarization, (Lin 04) showed that this statistic evaluation method achieves high correlation with human judgments.

### 3.2 Evaluation data set

The evaluation has been carried out using TED audio recordings, speeches from non-native English speakers presenting academic papers for approximately 15 minutes each. For most of these presentations, the corresponding paper is also available. Additionally, there are manual transcripts for 39 of the presentations, 37 of which with the corresponding paper.

Initially, the title, keywords and abstracts of these papers were taken as summaries of presentations, to be used as a gold standard to assess the quality of automatic summaries. But after some preliminary experiments we decided to discard paper abstracts, as the comparison of longer summaries is almost impossible, due to the important differences between texts of extracts from presentations and paper abstracts (discourse structure, vocabulary,...).

This problem also affects the comparison of 10- and 30-word automatic summaries with the title and keywords, but stylistic differences are less dramatic in shorter texts. That is why two additional gold standard *extractive* summaries were created for each presentation. Two human judges summarized the presentations by extracting chunks of words from the manual transcripts. For each document, summary models of about 10 (title or keywords style) and 30 words were manually produced by chunk extraction.

The voice files of 93 TED speeches, with paper summary model, were automatically transcribed using JANUS (Lavie *et al.* 94), a speaker independent ASR with a Word Error Rate (WER) of 31%. The ASR was trained with 31 of the 39 manual transcripts.

### 3.3 Evaluation conditions

To extract the collocation list, we have used the EuroSpeech'93 conference papers, trying to obtain domain-specific collocations.

Since oral input is ill-formed and the scope of the discourse markers cannot be calculated in terms of linguistic units, the scope has been stipulated as a 5-word window following it or a 10-word window around it, depending on the syntactic type of the marker.

In the PreSeg approach we established a size for the *Transcript Segmentation* module of 20 words per segment, similar to the average sentence length in the transcript corpus (20.81), while in the PostSeg *Extraction, Ranking and Selection* module, the chunk size limits ranged from 10 to 15, imitating the studied behavior of human summarizers, who usually extract shorter fragments to make summaries.

In all the experiments, lexical chains were computed taking into account only common nouns as candidate chain members. Due to the lack of punctuation and capitalization, usual textual methods to detect named entities are useless. In order to have a fair comparison between manual and automatic transcripts, capitalization was removed from the former. Only repetitions and synonyms have been considered, disregarding other Wordnet relations. In this domain, very general words (e.g., *speech or speed*) are used with a very specific sense, but since no domain-specific word-sense disambiguation is performed, considering all their Wordnet relations would yield an inadequate representation of the text.

As one of the model summaries to be compared with automatic summaries is a list of keywords without any particular order, we used the ROUGE-1 measure of unigram overlap as an evaluation of content overlap between summaries, with the following parameters: (i) 95 per cent confidence interval; (ii) stemming; (iii) stop words are not included in the calculations; (iv) a length limit is imposed.

| Summary Length | 10-word | | | 30-word | | |
|---|---|---|---|---|---|---|
| Summary references | **paper** T+K | **human** H1+H2 | **paperhuman** T+ K+H1+H2 | **paper** T+K | **human** H1+H2 | **paperhuman** T+K+H1+H2 |
| **manual transcriptions (37 docs)** | | | | | | |
| *Human 1* | *44.87* | – | – | *55.26* | – | – |
| *Human 2* | *37.38* | – | – | *50.12* | – | – |
| NFreq | 16.66 | 18.18 | 17.85 | 23.28 | 21.84 | 22.54 |
| NFirst | 19.88 | 26.67 | 24.40 | 34.76 | 20.92 | 24.92 |
| **ManSeg** | 26.39 | 34.59 | 31.06 | 41.07 | 29.90 | 32.75 |
| **PreSeg** | 24.99 | 32.11 | 29.04 | 38.67 | 33.66 | 35.25 |
| **PostSeg** | 26.16 | 31.65 | 29.20 | 39.61 | 32.04 | 34.13 |
| **automatic transcriptions (93 docs)** | | | | | | |
| NFreq | 18.61 | – | – | 18.46 | – | – |
| NFirst | 21.83 | – | – | 28.70 | – | – |
| **PreSeg** | 25.76 | – | – | 38.55 | – | – |
| **PostSeg** | 24.48 | – | – | 38.45 | – | – |

Table 2: Upper bound (*human ideal automatic systems*) and system performances when summarizing manual or automatic transcripts. ROUGE unigram overlap measure have been computed when taking, extract-based human summaries (H1:human1 and H2:human2), abstract-based author paper summaries (T:title and K:list of keywords) or all of them as model summaries.

## 4 Experiments and Results

The performance of the summarizers has been evaluated at two levels, producing summaries of different lengths (10 or 30 words) and having manual or automatic transcripts as input. In order to establish an upper bound for extract-based summarizers, the summaries produced by humans were evaluated as if they were the output of two ideal systems. Two kinds of baseline summaries were also created: the first $n$ (10 or 30) words of each transcript and the $n$-long segment maximizing the frequency of the words contained in it.

Table 2 shows the results of the experiments when comparing extract-based summaries from scientific presentation transcripts against three sets of models: the title and keywords of the corresponding paper (*paper*), both human summary models (*human*) and all of them (*paper + human*). Three different approaches to segmentation are detailed: besides the PreSeg and PostSeg approaches, a third ManSeg approach was included, with input manually segmented at sentence level.

### 4.1 Analysis of the results

As could be expected, recovering the content of model summaries in 10 words is more difficult than in 30 words, as seen in Table 2.

It is noteworthy that, in 10-word summaries, automatic summaries are more similar to human summaries than to the title and keywords of the paper. However, this tendency is reversed for 30-word summaries.

To analyze the obtained scores and determine the agreement between the different models, we have studied the correlation between them. Taking as a reference one model at a time, we have computed the ROUGE-1 measures for each kind of automatic summary. After that, the level of agreement between pairs of models has been measured comparing the values given to each summary according to the models. The Pearson correlation coefficient has been used to quantify this agreement, reaching a 99% confidence level in all cases, which indicates that different models agree significantly. Table 3 displays correlation values for each pair of models.

For 10-word summaries from manual tran-

| Manual-10 | Title | Human1 | Human2 |
|---|---|---|---|
| KeyWords | 0.567 | 0.529 | 0.495 |
| Title | | 0.646 | 0.633 |
| Human1 | | | 0.744 |

| Manual-30 | Title | Human1 | Human2 |
|---|---|---|---|
| KeyWords | 0.457 | 0.493 | 0.397 |
| Title | | 0.565 | 0.340 |
| Human1 | | | 0.608 |

| Automatic-10 | Title |
|---|---|
| KeyWords | 0.661 |

| Automatic-30 | Title |
|---|---|
| KeyWords | 0.679 |

Table 3: Pearson correlation coefficients between the different models in each evaluation set.

scripts, one of the humans correlates slightly better with paper models (KeyWords and Title) (0.53 and 0.65, respectively) than the other human (0.50 and 0.63). The same tendency is accentuated in 30-word summaries: 0.49 and 0.56 for the first judge, 0.40 and 0.34 for the second. That explains the fact that, when compared to paper-based summaries (see Table 2), Human2 obtains lower performance than Human1 (37.38, 50.12 *vs.* 44.87, 55.26, respectively) The style of human extracts seems closer to the titles of the papers than to the list of keywords, although differences between different kinds of paper-based summaries and human extracts are smaller in 30-word summaries. In any case, it is interesting to note that the correlation between Human1 and Human2 is the highest of all in both summary sizes (0.74 for 10-word, 0.61 for 30-word). The correlation between the KeyWords and Title models is also significant, especially when evaluating automatic transcripts (0.66 and 0.65).

A t-student test with 95% confidence interval was applied to determine whether differences between the summaries produced by the different systems (detailed in Table 2) were significant. When summarizing manual transcripts, humans always perform significantly better than any automatic system, as compared to paper-based summaries. For 10-word summaries, H1 summarizes better than H2. Differences between ManSeg, PostSeg and PreSeg are never significant, but the NFreq baseline is clearly the worst automatic system, and Nfirst is worse than ManSeg.

On the other hand, when having as input automatic transcripts, differences between PostSeg and PreSeg are not significative. While for 30-word summaries PreSeg and PostSeg perform significantly better than any baseline, for 10-word summaries the only significant differences are that NFreq is the worst system and PreSeg is better than Nfirst.

Finally, it is interesting to note that the quality of summaries does not drop when automatic, instead of manual, transcripts are used. This shows that the proposed techniques are robust enough to exploit salient textual features that allow to identify and delimit the most relevant parts of a presentation. Table 4 shows the 10-word output of the different systems, the summaries used as refences (KeyWords and Title), as well as the summaries produced by human judges.

## 5 Conclusions

We have presented two lexical chain based summarizer prototypes for spontaneous oral presentations of scientific papers. These prototypes rely on robust summarization techniques that have been successfully applied before to written text summarization, and show that they are robust enough to identify and delimit the most relevant parts of a document, even when the input text is linguistically ill-structured, as is proper of spontaneous speech, and contains transcription errors due to the low performance of automatic speech recognition (weighted error rate 31%).

Two different approaches to the detection of units of content have been tested, one segmenting the text in the preprocess step and the other exploiting the distribution of content as an additional feature to detect summary units. The performance obtained is ac-

| System Summary | |
|---|---|
| NFreq | generated pitch period and fluctuation a pitch period periodic repetition |
| NFirst | thank you and i good afternoon i messages fell into |
| **PreSeg** | about the new excitation model for the distance speech synthesis |
| **PostSeg** | excitation model for the distance speech synthesis quality of text |
| **Summary Reference** | |
| KeyWords | excitation vocoder text-to-speech synthesis |
| Title | a new model of excitation for text-to-speech synthesis |
| *Human 1* | distance speech synthesis text to speech synthesis excitation model residual signal |
| *Human 2* | new excitation model for the distance speech synthesis residual signal |

Table 4: Sample of 10-word summaries produced from an automatic transcript, the corresponding title and keywords used as reference and summaries produced by human judges.

ceptable in both cases.

We have shown that these techniques applied to summarizing transcribed scientific presentations produce results similar to human-produced summaries of the manual transcripts of these talks. Moreover, they are also similar enough to the title and keywords of the corresponding papers. Correlation between models has been studied, showing that the title and the list of keywords are good summary models to evaluate systems to summarize scientific presentations, although judge-made summaries are more stable.

A dummy baseline approach consisting of taking the first part of the document as summary does very poorly, in contrast with the performance of this same baseline in other genres, like newspaper articles. A baseline consisting of the segment that maximizes the frequency of words it contains performs even worse.

Finally, we have shown that the quality of summaries does not drop when automatic transcripts are used, instead of manual ones, so the approach is resistant to the noisy input of automatic speech recognizers.

## 6 Acknowledgements

This research has been partially funded by the European Commission project CHIL (IST-2004506969), the Spanish Research Department project ALIADO (TIC2002-04447), and by DURSI, the Research Department of the Catalan Government. TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI.

## References

(Alonso Alemany 05) Laura Alonso Alemany. Representing discourse for automatic text summarization via shallow NLP techniques. PhD thesis, Universitat de Barcelona, 2005.

(Arranz *et al.* 05) Victoria Arranz, Jordi Atserias, and Mauro Castillo. Multiwords and word sense disambiguation. In Proc. of CICLING 2005, Mexico City, 2005.

(Fellbaum 98) Christiane Fellbaum, editor. WordNet. An Electronic Lexical Database. Language, Speech, and Communication. The MIT Press, 1998.

(Kikuchi *et al.* 03) T. Kikuchi, S. Furui, and C. Hori. Automatic speech summarization based on sentence extraction and compaction. In ICASSP, volume I, pp 384–387, Hong Kong, 2003.

(Koumpis & Renals 00) K. Koumpis and S. Renals. Transcription and summarization of voicemail speech. In ICSLP, pp 688–691, 2000.

(Lavie *et al.* 94) A. Lavie, D. Gates, N. Coccaro, and L. Levin. Input segmentation of spontaneous speech in JANUS: a speech-to-speech translation system. In Proc. of the ANLP, 1994.

(Lin 04) Chin-Yew Lin. Looking for a few good metrics: Rouge and its evaluation. In Proc. of the NTCIR Workshop 4, Japan, 2004.

(Manning & Schütze 99) Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.

(Over & Yen 03) P. Over and J. Yen. An introduction to DUC 2003 - intrinsic evaluation of generic news text summarization systems. 2003.

(Stokes *et al.* 04) N. Stokes, E. Newman, J. Carthy, and A. F. Smeaton. Broadcast News Gisting using Lexical Cohesion Analysis. In Proc. of the 26th ECIR, pp. 209-222, Sunderland, U.K., 2004.