14

Radio Oranje: Enhanced Access to a Historical Spoken Word Collection

Laurens van der Werff, Willemijn Heeren, Roeland Ordelman, and Franciska de Jong
University of Twente

Abstract

Access to historical audio collections is typically very restricted: content is often only available on physical (analog) media and the metadata is usually limited to keywords, giving access at the level of relatively large fragments, e.g., an entire tape. Many spoken word heritage collections are now being digitized, which allows the introduction of more advanced search technology. This paper presents an approach that supports online access and search for recordings of historical speeches. A demonstrator has been built, based on the so-called Radio Oranje collection, which contains radio speeches by the Dutch Queen Wilhelmina that were broadcast during World War II. The audio has been aligned with its original 1940s manual transcriptions to create a time-stamped index that enables the speeches to be searched at the word level. Results are presented together with related photos from an external database.

Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands
Edited by: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde.

14.1 Introduction

At present, audio(visual) collections from the cultural heritage (CH) domain are at risk of becoming inaccessible, because (i) both the analog data carriers they are stored on are deteriorating and corresponding playback devices are becoming obsolete, and (ii) the materials are insufficiently disclosed for fast and easy access. In this paper we present a demonstrator for online access to a historical audio collection. The technical approach is based on a combination of speech processing and interaction design, and it has been applied to the collection of radio speeches that Queen Wilhelmina (1880-1962) addressed to the Dutch people during World War II – referred to as the 'Radio Oranje collection'. The speeches were broadcast via Radio Oranje, a radio channel set up in London, England, to inform the Dutch people in occupied areas. This demonstrator is an example of how indexing and access to audiovisual collections from the CH domain could be organized to overcome the limitations of traditional indexing methods for A/V material.

Preservation issues have been taken up in retrospective digitization projects for historic audio(visual) collections such as the EU IST PrestoSpace[1] project and the Dutch Beelden Voor De Toekomst[2]. In the case of the Radio Oranje collection, most recordings as well as their original 1940s transcripts underwent preservation measures and have recently been digitized by the Netherlands Institute for War Documentation (NIOD)[3] and the Netherlands Institute for Sound and Vision (NIBG) [4]. Without these measures, the Radio Oranje collection could only be accessed by reading the transcripts kept at the NIOD (in Amsterdam) and/or visiting the NIBG (in Hilversum) to obtain copies of the audio files. As collections become available digitally, they can be made accessible and, in principle, searchable via the Web.

To facilitate keyword search, some textual representation of the audiovisual documents is needed. For the kind of content under discussion here, descriptions typically consist of a set of keywords for long stretches of speech, e.g. an entire hour or tape. This type of metadata is not useless, but it is insufficiently specific to support all needs of a researcher: both the lack of precision in the description and the coarse time-resolution of retrieved results make the exploration of A/V documents quite cumbersome. Moreover, for most of the digitized and digital-born audiovisual documents, disclosure based on manual description is not an option, since manual annotation takes one to ten times the duration of a recording.

To improve access to digitized audiovisual collections it is therefore necessary to automatically generate time-stamped textual representations that describe the spoken content with much more precision (i.e. a higher time-resolution) than is the case in current practice. Automatic generation of a detailed index into the audio can be achieved in several ways, depending on the amount of metadata that is available for a collection. The extremes of the metadata dimension are a full

[1] http://www.prestospace.org/
[2] http://www.beeldenvoordetoekomst.nl
[3] http://www.niod.nl/
[4] http://www.beeldengeluid.nl/

manual transcript on one end, and no metadata at all at the other end. In the former case, aligning the transcription to the audio is sufficient for generating an index, in the latter case automatic speech recognition (ASR) can be employed for generating a textual representation of the spoken content.

In contrast to the broadcast news domain, which has been the main area of speech recognition research and benchmarks, speech from the CH domain can contain relatively large amounts of spontaneous speech (in which speakers overlap, hesitate, repeat themselves, etc.) and of speech that was recorded in adverse conditions (e.g. out on the street) or using suboptimal equipment. A number of research projects have aimed to advance ASR and spoken document retrieval specifically for the CH domain. In The National Gallery of the Spoken Word project, the SpeechFind spoken document retrieval system was developed: it automatically generates metadata for audio documents by segmenting the audio and generating ASR transcripts, and also makes the audio searchable through a Web interface (Hansen et al. 2005). The MALACH (Multilingual Access to Large spoken ArCHives) project investigated access to a vast collection of testimonies from Shoah survivors (Byrne et al. 2004). The goal of that project was to advance English and Czech ASR for the oral history domain and to study how recognition can be best incorporated in further processing and retrieval steps (Gustman et al. 2002). In the Netherlands, the CHoral project[5], part of the NWO-CATCH[6] program, investigates technology for indexing and accessing Dutch, historically relevant spoken documents (Ordelman et al. 2006).

In this paper we will describe a framework for improved access to spoken CH-content. More specifically, we will describe the steps taken to improve access to the Radio Oranje collection. In section 14.2 we will focus on the synchronization step, also called alignment, where the 1940s transcripts were used to generate a time-stamped index of the spoken documents. Section 14.3 discusses how this index was exploited to support online search and browsing and how it was used to enhance presentation. The generation of cross-links to present the speeches together with related photos from an external database will also be explained in this section. Remaining issues and future work are discussed in section 14.4.

14.2 Optimizing alignment

Given the poor sound quality of the speeches – the original recordings were made on historical equipment and contain hiss, pops, and scratches – an ASR engine would not be able to generate an adequate transcript. In the case of an alignment task, audio frames are linked to a phonetic representation of a manual transcript using acoustic models from an ASR system. Alignment is much more robust towards mismatches between models and data than ASR. An example of access to a video archive using alignment of manual transcripts can be found in Christel et al. (2006).

[5] http://hmi.ewi.utwente.nl/choral
[6] http://www.nwo.nl/catch

The collection of speeches in the Radio Oranje project have been fully transcribed during the war and therefore alignment could be done for this collection. The data under consideration consisted of 29 speeches by Queen Wilhelmina, with lengths varying between 5 and 19 minutes. All speeches were manually segmented at the sentence level, giving a total of 853 sentence-sized segments with an average length of 15.7 seconds. For evaluation purposes, two full speeches were segmented at the word level yielding 2028 manually aligned word boundaries. The alignment tool from an off-the-shelf multi-mixture Gaussian HMM-based speech recognition engine was used (Pellom 2001), which produces Viterbi optimized word-based alignments.

14.2.1 Experiment I: Acoustic models

In contrast to an ASR system, which generates a hypothesis of *what* was said, an alignment task only has to decide *where* something was said. Traditionally the same acoustic models are used for both alignment and recognition, but this need not automatically lead to the best alignment result.

We first performed an alignment using gender- and speaker-independent acoustic models, optimized for broadcast news (BN) (de Jong et al. 2006). Both triphone (context-dependent) and monophone (context-independent) BN models were used. New acoustic models were trained from the resulting alignments leading to a speaker-dependent acoustic model. This was then used to perform a second iteration for training the final Wilhelmina models. In total, three different acoustic models were evaluated: a triphone BN model, a monophone BN model and a monophone Wilhelmina model. For these experiments, sentence-sized segments were used as input and the resulting alignment was evaluated at the word level.

14.2.2 Experiment II: Segment size

An alignment tool assigns acoustic model states to each of the audio frames, based on the phonemes that are predicted from the transcription. This is done in such a way that the total likelihood of this state sequence, given the audio, is maximized. Due to the complexity of the task it is not feasible to exhaustively explore all possible alignments. In practice, some pruning is applied and the alignment will converge around a local optimum.

An anchor point is a mark in the audio and the transcription that ties two equivalent positions together. A segment can be viewed as the audio fragment between two adjacent anchor points. When more anchor points are provided to the alignment tool, the task of aligning becomes easier and pruning becomes less of an issue. To determine the influence of segment size on alignment quality, experiments were performed in which alignment was done on varying input sizes. The results were evaluated at the word level.

14.2.3 Experiment III: Grapheme-to-phoneme conversion

Alignment between text and speech is not done directly but through a phoneme representation of the text. First the orthographic transcription is converted into a phonetic representation and then a sequence of acoustic models corresponding to these phonemes can be aligned to the audio. The conversion of graphemes to phonemes has been extensively studied in the past, see Strik and Cucchiarini (1999) for a review. Most grapheme-to-phoneme (G2P) conversion tools produce a canonical phonetic transcription based on a background dictionary that is augmented with a rule-based system. Both the background lexicon and the rules are usually based on modern spelling and the corresponding current pronunciation.

In the CH domain, transcriptions can use archaic spelling conventions as was the case with this collection (e.g. *eisch* instead of *eis*, meaning 'demand', and *voorteekenen* instead of *voortekenen*, meaning 'omens'). To investigate the influence of the G2P conversion on alignment performance, three different phonetic versions of the reference texts were produced and compared: (i) a fully automatic G2P version, (ii) a G2P performed on a version of the reference texts after conversion to modern spelling conventions, and (iii) a manually checked phonetic conversion (thus excluding automatic G2P errors from the process).

14.2.4 Results

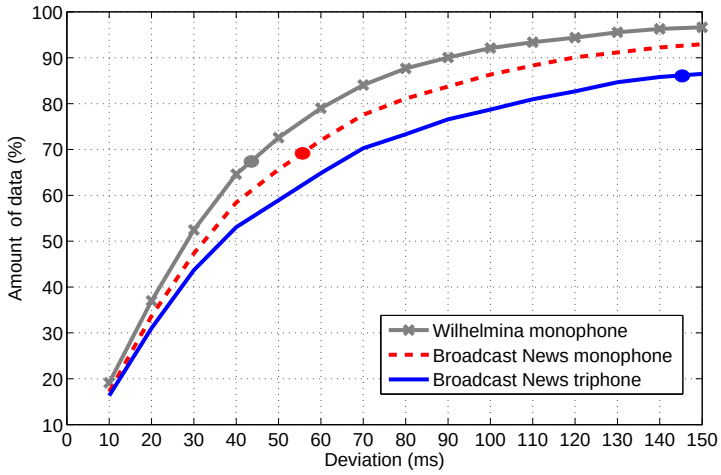

Figure 14.1: Acoustic model performance. For each of the three acoustic models the amount of data complying with a certain amount of deviation from the reference transcript is shown.

Figure 14.1 shows the percentage of word boundaries (vertical axis) that fall

within a certain deviation from the manual reference alignment (horizontal axis). The dots mark the average deviation from the reference. When considering this average deviation, BN monophone acoustic models performed nearly 60% better than traditional BN triphone models on this task. Acoustic models that were specifically trained on these speeches provided an added improvement of almost 20%. The maximum deviation from the reference for all monophone models was less than one second. Regardless of the performance level required, monophone models scored better than triphone models and data-matched models scored better than generic BN models.

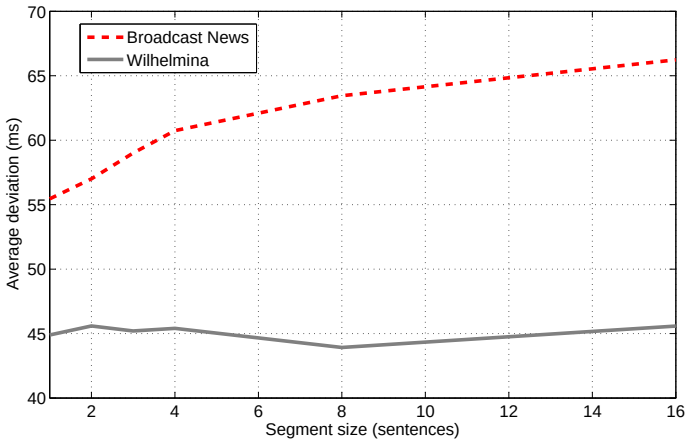

Figure 14.2: Alignment performance as a function of segment size for the BN triphone models and the speaker-adapted Wilhelmina models.

The performance figures that were found for this set are lower than those found in previous studies, see for example Brugnara et al. (1993), where 89% of the aligned phonemes were found within 20 ms of the reference. This stresses the mismatch that exists between the generic BN acoustic models and the historic audio under consideration. Another reason for this difference is that evaluation was done for word-boundaries only, not for phoneme boundaries. This affected the accuracy of the manual placement of reference boundaries, but – as was found in for example Rapp (1995) – it also leads to a slight reduction in overall alignment performance.

Figure 14.2 shows the average alignment error for varying segment sizes. When the data-matched Wilhelmina model is used, segment size seems relatively uncritical. Segments with a length of up to around five minutes do not show a significant reduction in alignment performance when compared to the original sentence-sized segments. Aligning long segments with BN triphone models re-

quired a reduction in pruning that led to a high increase in processing time (>10 times longer).

	Phones altered (%)	Average deviation (ms)
Original spelling	0	55
Modern spelling	1	56
Manual conversion	5	54

Table 14.1: The effect of grapheme-to-phoneme conversion method on alignment performance.

Table 14.1 gives the results for the three types of G2P when the BN monophone acoustic models are used. Not only is the impact of old spelling conventions on G2P quite limited (only 1% of all phonemes is affected), the differences that do exist turn out to be of no consequence for finding the word boundaries. Removing all grapheme-to-phoneme conversion errors from the transcription also shows no significant improvement on alignment performance at the level of word boundaries.

14.2.5 Summary

Overall, alignment performance was more than adequate for this task. The duration of an average syllable lies in the range of 100-300 ms and over 90% of all detected word boundaries were found within 100 ms of the reference. The use of monophone models resulted in better alignment performance than use of traditional triphone models. Segment size was relatively uncritical when the models were well matched to the data. In the case that mismatch between the audio and the models was high, much more processing time was required to obtain acceptable alignment results. Finally, despite the 1940s spelling conventions, there was no impact of grapheme-to-phoneme conversion errors on locating word boundaries.

14.3 Radio Oranje Web interface

On the basis of the alignment, a time-stamped index was created that allows word-level access to the speeches through the Radio Oranje Web interface[7]. The index also facilitated development of additional functionalities for user support. The user experience was enhanced through the generation of cross-links to a topically-related photo collection.

[7]http://hmi.ewi.utwente.nl/choral/radiooranje.html (in Dutch)

14.3.1 Accessing the spoken word documents

For search and browsing, the interface allows entry to the collection at two levels: an entire speech or a speech fragment. It is expected that users will enter their queries in contemporary Dutch spelling, whereas the index contains Dutch in the 1940s spelling. To prevent that words written in the old-fashioned spelling become irretrievable, a dictionary was used to translate terms from user queries into index terms. This dictionary was created manually given the relatively small scale of the task. Boolean retrieval is currently supported and query results are ranked by date, showing the speech's title, broadcast date and duration as well as an excerpt of the relevant sentence fragment. If the framework is used for larger CH collections, more advanced (ranked) retrieval techniques should be used.

Once the user selects a particular spoken word document, basic playback options (start-stop-pause) are insufficient for navigation, as linear examination of the fragments from the result list is relatively inefficient. Therefore, more elaborate and dynamic user controls and content visualization options are needed. In earlier research, visual content representations have been developed that for instance indicate speaker turns (e.g. Slaughter et al. (1998)) or the occurrence(s) of query terms in time (e.g. Whittaker et al. (1999), Christel et al. (2006)). Other tools developed for faster browsing allow users to speed up audio playback, since time-compressed speech remains intelligible up to double its original speed, e.g. Hürst et al. (2004).

To offer a proper mix of flexibility and transparency we developed an interactive visualization of the audio content on the basis of the time-stamped index. It shows an overview of the entire speech as well as a zoomed-in view of a 45 s window around the cursor. The exact positions of highlights, e.g. query terms and sentence boundaries, are shown in both bars. Through this combination of bars it gives a clear overview of the document as well as detailed information on the fragment that is currently being played. This combination is new. Furthermore, the visualization is interactive: clicking any point on either bar will restart the audio at that point in time, which allows the user to quickly browse through the spoken document.

During audio playback, users prefer to take control of playback over predetermined play durations, since restricted playback may stop at unpredictable places (Whittaker et al. 1998). Another issue that may be encountered during playback is that query terms occur right at the beginning of the retrieved fragment. The relevant term may be played before users are well-aware of it. In the W.F. Hermans system[8] this problem was overcome by enabling the user to select the size of the fragment's context (Huijbregts et al. 2005). In the current interface we chose to add an extra button for restarting the fragment from the original entry point.

The second functionality that was added to support users during playback was subtitling. This highlights the word being spoken and shows the query terms in a contrasting color. Subtitling was added to aid intelligibility given the sometimes poor audio quality and the old-fashioned, formal language use encountered in the speeches of Queen Wilhelmina.

[8] http://www.willemfrederikhermans.nl/multimedia/

14.3.2 Cross-media linking

In the CH domain, the ongoing digitization of historical texts, images, pamphlets, photos, audiovisual materials etc. makes it possible to (i) automatically identify links between documents from a variety of modalities and/or collections and (ii) present related documents in one multimedia presentation. These possibilities create new opportunities for comparative research in for instance the historical domain and for the presentation of documents from audiovisual archives for educational purposes. In cross-media linking, content from different media types is associated. This is done by linking the semantic representations from each media type either directly or through, for instance, a thesaurus or ontology. An example is the cross-media browser Infolink that combines broadcast news videos with data from a historical video archive and textual information from a newspaper corpus (Morang et al. 2005).

In our demonstrator, spoken word fragments and photographic material on the same topic, i.e. World War II, were linked. The photographic material was taken from a collection of over 55,000 photos maintained by the NIOD: the photos are partly from the same period as the Radio Oranje broadcasts. However, since unrestricted access to the photo database with elaborate descriptions was not obtained, fully automatic search could not be investigated. The restricted metadata that was available consisted of a few keywords per photo. These keywords were automatically extracted from the online catalog for the photo collection.

Searching and browsing functionality were developed for the spoken content, and as a pre-processing step sets of photos were semi-automatically linked to the speeches. Since the Radio Oranje collection is characterized by a very formal and metaphorical speaking style, it was not possible to automatically match the spoken content to the keywords from the photo database. Therefore, semantic representations for the speeches were generated by manually assigning one or more keywords from the photo thesaurus to each speech. This was done on the basis of its title and global content. The 29 speeches were described by (combinations) of 18 keywords such as Liberation, May 1940, Christmas or Netherlands Indies. These keywords defined photo sets ranging in size from 2 to 200. A number of keywords that were relevant to the entire collection was selected as a default set of photos when speech-level sets are too small, for example: Illegal Press & Radio, Queen Wilhelmina and Dutch Street Scenes.

While the audio is being played, photos that relate to the topic are shown with a refresh rate of ten seconds. Figure 14.3 shows the resulting multimedia presentation: during audio playback the information visualization, subtitling and topically-related photos are shown.

14.4 Discussion and future work

In this paper we have presented the Radio Oranje demonstrator, which is an instantiation of the framework for enhanced spoken word access developed as part of the CHoral project. It shows how access to audiovisual databases from the CH do-

Figure 14.3: Screen shot of the playback interface showing a related photo, subtitling and the interactive browsing bars.

main can be changed using currently available technology for automatic indexing, information retrieval and content visualization.

With respect to the alignment results presented in section 14.2.4, the improved performance from using monophone vs triphone models was in line with expectations (van Santen and Sproat 1999). Although there are some techniques available to improve alignment, such as systematic bias removal (Dines et al. 2002) or spectral boundary correction (Kim and Concie 2002), these were not deemed necessary for the development of this particular system.

To support users during audio browsing and to make their searches as efficient and satisfactory as possible, we developed the information visualization component presented in section 14.3.1. It enables the user to quickly estimate the location of the most important regions within a document given his/her query. In the present demonstrator system, those regions truly contain the query terms given the accurate transcripts. Even if the transcripts were not fully accurate (due to ASR errors for example), the user is expected to be much faster in judging a fragment's relevance using this visualization than without any information on the location of highlights or with less specific information visualizations, see e.g. Whittaker et al. (1999) and Hearst (1995). Future work in the CHoral project will determine how users can be supported even better during retrieval of historical spoken documents.

Another issue for future research concerns semantics. The semantic gap, i.e. the fact that the match between the words spoken and the topic that is being talked about is only partial, should be investigated further. Since manual annotation of high-level semantic information is too (time-)costly, automatic extraction might be a feasible approach. Keywords should ideally be limited to a controlled vocabulary to enable cross-linking with other collections and media. Mapping the terms in the transcription to this vocabulary can be done using a thesaurus- or ontology-type approach as in Wordnet (Fellbaum 1998). Moreover, audiovisual documents on

specific periods or events in history – such as World War II – require the addition of expert knowledge for successfully matching user queries. Words get specific connotations in the context of certain historical periods or events (e.g., euphemisms) that cannot be solved by standard solutions such as document or query expansion using synonyms, hyponyms and hypernyms. Such mappings can – for now – only be provided through manual effort.

In sum, the framework for enhanced spoken word access will be developed further within the CHoral project in order to enable widespread use of Dutch historical spoken word documents in research, education and content production.

Acknowledgements

The research reported on here was funded by the NWO project CHoral, part of CATCH, and supported by the research program MultimediaN (http://www.multimedian.nl). MultimediaN is sponsored by the Dutch government under contract BSIK 03-31.

References

Brugnara, F., Falavigna, D. and Omologo, M.(1993), Automatic segmentation and labeling of speech based on Hidden Markov Models, *Speech Communication* **12**(4), 357–370.

Byrne, W., D.Doermann, Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T. and Zhu, W.-J.(2004), Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives, *IEEE Trans. Speech Audio Proc.*

Christel, M., Richardson, J. and Wactlar, H.(2006), Facilitating access to large digital oral history archives through Informedia technologies, *Proceedings of JCDL '06*, pp. 194–195.

de Jong, F., Ordelman, R. and Huijbregts, M.(2006), Automated speech and audio analysis for semantic access to multimedia, *in* Y. Avrithis, Y. Kompatsiaris, S. Staab and N. O'Connor (eds), *Proceedings of the First International Conference on Semantic and Digital Media Technologies, SAMT 2006*, Vol. 4306 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 226–240. ISBN=3-540-49335-2.

Dines, J., Sridharan, S. and Moody, M.(2002), Automatic speech segmentation with hmm, *Proceedings of the 9th Australian Conference on Speech Science and Technology*.

Fellbaum, C. (ed.)(1998), *Wordnet. An electronic lexical database*, MIT Press, Cambridge, MA.

Gustman, S., Soergel, D., Oard, D., Byrne, W., Picheny, M., Ramabhadran, B. and Greenberg, D.(2002), Supporting Access to Large Digital Oral History Archives, *Proceedings of the Joint Conference on Digital Libraries*, pp. 18–27.

Hansen, J., Huang, R., Zhou, B., Deadle, M., Deller, J., Gurijala, A., Kurimo, M. and Angkititrakul, P.(2005), SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word, *IEEE Transactions on Speech and Audio Processing* **13**(5), 712–730.

Hearst, M. A.(1995), TileBars: Visualization of Term Distribution Information in Full Text Information Access, *Proceedings of the Conference on Human Factors in Computing Systems, CHI'95*.

Huijbregts, M., Ordelman, R. and de Jong, F.(2005), A Spoken Document Retrieval Application in the Oral History Domain, *Proceedings of 10th international conference Speech and Computer, Patras, Greece (SPECOM 2005)*, 2, University of Patras, Wire Communications Laboratory Moscow State Linguistics University, pp. 699–702. ISBN=5-7452-0110-X.

Hürst, W., Lauer, T. and Götz, G.(2004), An elastic audio slider for interactive speech skimming, *Proceedings of NordCHI '04*.

Kim, Y.-J. and Concie, A.(2002), Automatic segmentation combining an hmm-based approach and spectral boundary correction, *Proceedings of ICSLP 2002*, pp. 145–148.

Morang, J., Ordelman, R., de Jong, F. and van Hessen, A.(2005), Infolink: analysis of Dutch broadcast news and cross-media browsing, *Proceedings of IEEE International Conference on Multimedia and Expo, ICME 2005*, pp. 1582–1585.

Ordelman, R., de Jong, F. and Heeren, W.(2006), Exploration of Audiovisual Heritage Using Audio Indexing Technology, *Proc. of the 1st workshop on Intelligent Technologies for Cultural Heritage Exploitation*, pp. 36–39.

Pellom, B.(2001), SONIC: The University of Colorado Continuous Speech Recognizer. Technical Report TR-CSLR-2001-01, University of Colorado.

Rapp, S.(1995), Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German, *Proceedings of ELSNET goes east and IMACS Workshop "Integration of Language and Speech in Academia and Industry"*.

Slaughter, L., Oard, D. W., Warnick, V. L., Harding, J. L. and Wilkerson, G. J.(1998), A graphical Interface for Speech-Based Retrieval, *ACM DL*, pp. 305–306.

Strik, H. and Cucchiarini, C.(1999), Modeling pronunciation variation for ASR: A survey of the literature, *Speech Communication* **29**, 225–246.

van Santen, J. and Sproat, R.(1999), High-accuracy automatic segmentation, *Proceedings of EuroSpeech99*.

Whittaker, S., Hirschberg, J. and Nakatani, C.(1998), Play it again: a study of the factors underlying speech browsing behavior, *Proceedings of CHI 1998*.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F. C. N. and Singhal, A.(1999), SCAN: Designing and Evaluating User Interfaces to Support Retrieval From Speech Archives, *Research and Development in Information Retrieval*, pp. 26–33.