Detecting and Labeling Folk Literature in Spoken Cultural Heritage Archives using Structural and Prosodic Features

Fabio Valente, Petr Motlicek
Idiap Research Institute, CH-1920 Martigny, Switzerland
{fabio.valente,petr.motlicek}@idiap.ch

Abstract

Spoken cultural heritage can present considerably heterogeneous content as tales, stories, recitals, poems, theatrical representations and other form of folk literature. This work investigates the automatic detection and classification of those data type in large spoken audio archives. The corpus used for this study consists of 90 radio broadcast shows collected for preserving a large variety of Swiss French dialects. Given the variability of the language spoken in the recordings, the paper proposes a language-independent system based on structural features obtained using a speaker diarization system and various acoustic/prosodic features. Results reveal that such a system can achieve an F-measure equal to 0.85 (Precision 0.88/ Recall 0.84) in retrieving folk literature in those archives. Prosodic features appear more effective and complementary to structural features. Furthermore, the paper investigates whether the same approach can be used to label speech segments into five large classes (Storytelling, Poetry, Theatre, Interviews, Functionals) showing F-measures ranging from 0.52 to 0.88. As last contribution, prosodic features for disambiguating between spoken prose and spoken poetry are investigated. In summary the study shows that simple structural and acoustic/prosodic features can be used to effectively retrieve and label folk literature in broadcast archives.

1. Introduction

Audio archives of cultural heritage represent an important form of saving people's collective memories. Given the size and the complexity of those archives, audio and speech processing technologies have been applied to automatically structure, index and access those data [6]. For instance, efforts like the SpeechFind project [7], the MALACH project [5] and the CHoral project [12] made use of various speech processing and information retrieval techniques to improve access to testimonies and interviews on historical events. Beside testimonies and interviews, spoken data from the cultural heritage also includes folk literature, i.e., oral traditions of cultures having no written form. Example of folk literature includes tales, stories, recitals, poems and theatrical representations. As universal, most countries devoted collection campaigns to preserve this spoken heritage. This work investigates whether folk literature can be automatically detected and labeled in large unstructured spoken audio archives.

The data used for this study consists of a subset of radio broadcast shows belonging to the collection *Archives des parlers patois de la Suisse romande et des regions voisines*[1] from the Radio Suisse Romande. The radio shows have been broadcasted between 1950 and 1980 and are devoted to the preservation and the disclosure of Swiss French dialects from various regions (Fribourg, Valais, Vaud, Jura) as well as neighboring regions from Italy and France. Beside more conventional contents like interviews, presentations and discussions, almost half of the broadcasts consists of folk literature (stories, poems, theatrical representations, tales) spoken in various dialects. Automatically processing such a dataset presents various challenges including the different quality of the recordings spanning a time-line of over 30 years and the difficulties in obtaining reliable speech transcript through Automatic Speech Recognition. In fact, the languages spoken are often local dialects where very little training data is available.

Most of the previous efforts in the area of spoken cultural heritage has focused on data like interviews or testimonies [5], [12]. This paper investigates whether folk literature can be automatically detected and labeled. In order to overcome problems related to dialect/language data sparseness, this work investigates the use of language independent features that could be easily and robustly estimated, i.e., prosodic/stylistic features and structural features extracted through a speaker diarization system. The remainder of the paper is organized as follows: Section 2 describes the data setup used in the study as well as the taxonomy used to label the different type of folk literature, Section 3 describes the data processing based on speaker diarization and the struc-

[1] http://son.memovs.ch/S024/doc/page_patois.htm

Story Label	Story Structure	Speakers	Speaking style
Functionals	Monologues	Professional	Prompted/Scripted
Interview	Dialogues/Multiparty	Professional/non-Professional	Spontaneous
Storytelling	Monologues	Professional/non-Professional	Narrative/Expressive
Poetry	Monologues	Professional/non-Professional	Narrative/Expressive
Theatre	Monologues/Dialogues/Multiparty	Professional	Expressive

Table 1. Summary of structural, stylistic and speaker properties for the different story labels. The speaking style properties follows the descriptions defined in [9].

tural/prosodic feature extraction, Section 4 describes the experiments and the paper is concluded in Section 5.

Story Label	Number of Stories	Time Percentage
Storytelling	92	16%
Poetry	67	10%
Theatre	26	15%
Functionals	230	16%
Interview	128	35%
Others	24	8%

Table 2. Story distributions in the corpus including the number of stories and the time percentage.

2. Data Description

The data used for this study consists of radio broadcast shows belonging to the collection *Archives des parlers patois de la Suisse romande et des regions voisines* from the Radio Suisse Romande. The radio shows have been broadcasted between 1950 and 1980 and are devoted to the preservation and disclosure of Swiss French dialects from various regions (Fribourg, Valais, Vaud, Jura) as well as neighboring regions from Italy and France. The corpus includes over 1500 recordings; a small subset of them also includes the show scripts with the approximated "story" boundaries[2] as well as a short description of the different "stories", the dialect spoken, the name of participants and their roles (anchorman, guests, actors). 90 shows were uniformly sampled across the 30 years thus considering very different acoustic quality in the audio. The 90 recordings have an average duration of 25 minutes accounting for approximatively 35 hours of speech. The program scripts are available for these 90 shows. The radio broadcast shows contain 60 different story labels including interviews, tales, poetry, recitals, theatrical representations, discussions, bibliographical references and so on. Precise story boundaries are obtained manually aligning the broadcast show scripts on the audio tracks thus generating precise start and end time for each of the different story segments. After that, the story labels from the show scripts are clustered together in **folk literature** (tales, history, legends, recitals, anecdotes, poems, theatre) and **conventional broadcast data** (introductions, openings, interview, debates, comments).

To study a finer classification scheme, also the following six classes obtained considering stylistic and structural differences between stories are considered :

1 **Storytelling**: includes a wide set of labels like tales, history, legends, recitals, anecdotes.
2 **Poetry**: includes labels like poems, poetry and rimes.
3 **Theatre**: includes theatrical representations.
4 **Functional**: includes introductions, presentations, conclusions, bibliographical notes, comments and other labels related with the functioning of the broadcast show and are typically spoken by professional speakers, e.g., the show anchormen.
5 **Interview**: includes speech segments where a professional speaker, e.g., the anchorman or a journalist interviews one or more guests.
6 **Other**: includes the remaining labels which do not belong to any of the previous categories. This class is included as garbage model to avoid training and testing on a very small and sparse classes.

While functionals are typically spoken in French, the other story types are typically spoken in different dialects. The last three categories are conventional story types widely studied in audio data like broadcast news; the first three are instances of folk literature and represent an important portion of those archives. Statistics of the six different story labels are reported in Table 2 for the 90 recordings. It can be noticed that almost 40% of the audio consists of folk literature. Beside obvious differences in content, the six categories have also differences related to their structure and speaking styles. For instance functionals, storytelling and poetry are monologues characterized by long speaker turns while interviews and theatrical representations are typically dialogues or multi-party conversations between two or more speakers. Functionals are typically uttered by professional speakers, i.e., the show anchormen, which exhibit a prompted/scripted speaking style (see [9] for a speaking style review) while introducing the stories, presenting guests or commenting. Interviews are dialogues between a professional speaker, e.g., a journalist and a non-professional one, e.g., an interviewee, which exhibit a spontaneous speaking style [9]. On the other hand, storytelling, poetry and theatrical representations can be uttered by both

[2]The term stories in this work refers to semantically uniform audio segments.

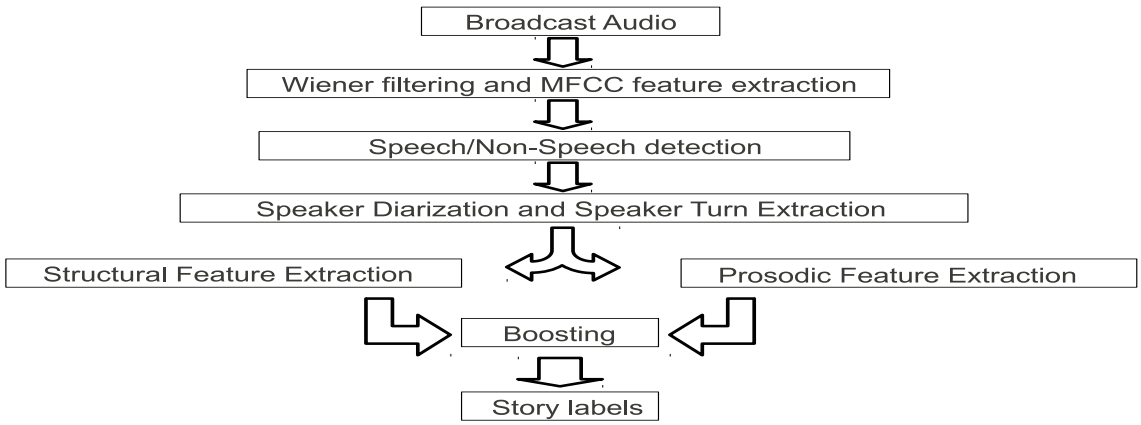

Figure 1. Schematic representation of the labeling system.

professional, e.g., actors, or non-professional speakers and they are characterized by a narrative/expressive speaking styles [18]. Beside expressiveness, poetry and poems are also characterized by meters (recurrent timing and beating patterns) [11, 8]. Those structural and stylistic differences in between stories are summarized in Table 1.

Many previous works made use of those differences to segment broadcast news data into topics and stories [16, 13], recognize speaker roles [3, 4] or summarize the content [10]. This work investigates whether the same structural and prosodic differences can be useful to automatically detect and classify spoken folk literature in cultural heritage audio archives.

3 System Overview

The detection and the labeling are based on a supervised approach in which a boosting algorithm trained on structural and prosodic features assigns to each speaker turn a label in between those in Table 1. The overall system is schematically depicted in Figure 1. The processing starts by signal de-noising and acoustic feature extraction (MFCC) followed by a speech/non-speech classification. The broadcast speech regions are then used as input for a speaker diarization system which infers "who spoke when" in the audio file performing two simultaneous tasks: inferring the number of speakers in the show and assigning each speech segment to one of them. The diarization output is then used to extract a sequence of *speaker turns* (see Section 3). Several *stylistic and structural features* are then extracted on a turn basis (see Sections 3.2 and 3.3) . Given the heterogeneity of those various features (both discrete and continuous), a *booster classifier* is used (see Section 3.4) to combine them.

In a first task, the booster classifier is trained to recognize folk spoken literature versus conventional broadcast. In a second task, the classifier is trained to assign to each turn a label in between those described in section 2. In the following the various modules that compose the system are briefly described.

3.1 Automatic Speaker Turn Extraction

The raw audio is pre-processed using a Wiener filter denoising as described in [1] in order to reduce noisy artefacts; after that, 19 MFCC coefficients are extracted from 30ms windows shifted every 10ms.

Speech/ non-speech detection is performed using a three state ergodic Hidden Markov Model (HMM) where the first state represents speech, the second represents silence and the third represents noise/music. Emission probabilities are modeled using Gaussian Mixture Models (GMM), and each GMM has 32 components. Models are trained on Broadcast audio.

The speech regions from the audio file are then used into an HMM/GMM speaker diarization system in which each speaker is represented by an HMM state with GMM emission probability [2]. The diarization starts with a uniform linear segmentation of the input into a large number of clusters (speakers). Successively, at each step, a cluster pair is merged based on a distance measure like the BIC or its modified version [2]. The merging stops when all the BIC values are less than zero. After each merge, a Viterbi realignment of speaker boundaries is performed with the estimated speaker models. This system showed state-of-the-art performances in several recent NIST Rich Transcription evaluations.

Based on the speaker diarization output, a sequence of speaker turns is then extracted. We adopt a simplified definition of "turn" defined as a speech region from the same speaker uninterrupted by pauses longer then 500 ms while a "sentence" is a speech region from the same speaker uninterrupted by any silence/pause [16]. The speaker diarization produces an unique identifier for each speaker thus in order to exploit the patterns in which the speaker appears in the show (as for instance in [4]), the identifiers are sorted

so that s_1 corresponds to the first speaker appearing in the show, s_2 corresponds to the second speaker and so on. More formally, for each show the following triplets are available:

$$T = \{(t_1, \Delta t_1, I_1), ..., (t_N, \Delta t_N, I_N)\} \quad (1)$$

where t_n is the beginning time of the n-th turn, Δt_n is its duration, I_n is the speaker identifier associated with the turn. Based on the sequence in Eq.(1), structural and acoustic/prosodic feature sets are extracted from each turn.

3.2 Structural Feature Extraction

Structural features used in this study are similar to those described in [10]. They consists in **turn duration, the maximum and the minimum sentence duration in the turn, the number of sentences in the turn, the relative position of the turn in the show, the ratio between amount of speech and amount of silence in the turn, the speaker identifier associated with the turn** (under the rationale that the first speakers are typically the show anchormen). Those turn-based statistics do not capture longer term structure and patterns that could be useful to determine the type of story. For instance interviews can be considered a sequence of short turns from a journalist and longer turns from a guest answering questions, while tales and poetry can be considered a sequences of long turns from the same speaker. In order to include this type of information, we introduce **N-gram of consecutive structural features** computed from the following and preceding turns. Third order n-gram, i.e., trigrams are used in this work. Let us for instance consider the case of speaker identifiers N-gram: in case of monologues, only N-gram containing the same speaker identifier will be different from zeros while in case of conversations (interviews or theatre) only N-gram containing different speakers will be different from zero.

Before computing those N-gram, the continuous features are quantized into 16 bins of equal area under the normal distribution. An independent N-gram set is estimated for each of the seven features. In such a way also statistics from neighboring turns are included with the aim of modeling recurrent patterns like sequences of short-long turns from different speakers or sequences of long turns from the same speaker.

3.3 Acoustic/Prosodic Feature Extraction

The second set of features consists in acoustic/prosodic measures extracted over the turn duration. Also those features have been largely used for determining speaker roles and speaking styles in broadcast data. They include **the average speaking rate, the average articulation rate, various F0 statistics (mean, median, minimum, maximum, variance and slope), and minimum, maximum, and mean RMS energy** extracted using Praat. The statistics are computed based on pseudo-syllables estimation (see [14]). They are used both as raw features and after a speaker-based histogram normalization.

Those features capture speaking style for the current turn. As before, in order to capture also informations from preceding/following turns, **N-gram of consecutive prosodic features** (see [17]) are used. The continuous features are quantized into 16 bins of equal area under the normal distribution. After that, the discrete prosodic feature f_n is augmented with the features of following and preceding turn so that the sequence $\{f_{n-1}, f_n, f_{n+1}\}$ is obtained, *N-gram* counts are computed and used in the booster classifier. N-gram counts are estimated for each of the prosodic features. Those features are expected to capture changes in speaking styles (professional to spontaneous like in interviews), or steady segments like for instance poetry and stories.

3.4 Boosting classifier

In order to integrate continuous, discrete features as well as N-gram counts into the same classifier, a boosting approach is used. The principle of boosting is to combine many weak learning classifiers to produce a single accurate classifier. The algorithm generates weak classification rules by calling the weak learners repeatedly in series of rounds. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The version of Boosting algorithm used was multi-class Boosting defined in [15]. The weak learners are one-level decision trees. This algorithm provides a very simple and effective way to combine continuous features as well as discrete features.

4 Experiments

Experiments are based on a leave-one-out approach where one show from the dataset is used for testing and the remaining 24 are used for training/development. Results are reported in terms of F-measure (Precision/Recall) for the two broad classes: folk literature versus conventional broadcast classes. Precision and Recall are computed as percentage of time correctly labeled/retrieved in order to take into account possible mismatches between the manual boundaries and the speaker diarization output. Results are reported in Table 3 in case of structural and prosodic features as well as their combination. Prosodic features achieve an F-measure equal to 0.80 (Precision 0.84/ Recall 0.77) in retrieving folk literature outperforming structural features which achieve only an F-measure equal to 0.69. Their combination produces an F-measure equal to 0.83 (Precision 0.87/ Recall 0.80) thus they appear complementary.

In a second experiments, the classifier attempts to assign one of the six labels [Functionals/Interviews/Other/Storytellying/Poetry/Theatre]. Results

	Prosodic	Structural	Prosodic+Structural	Prosodic+Structural + Prediction Features
Folk literature	0.80 (0.84/0.77)	0.69 (0.73/0.66)	0.83 (0.87/0.80)	0.85 (0.88/0.84)
Conventional Broadcast	0.91 (0.89/0.93)	0.87 (0.89/0.85)	0.92 (0.94/0.90)	0.93 (0.93/0.94)

Table 3. **F-measure (Precision/Recall) for detecting and labeling folk spoken data. Results are reported for prosodic and structural alone and in combination; the last column reports results whenever linear prediction features are also included.**

	Prosodic	Structural	Prosodic+Structural	Prosodic+Structural + Prediction Features
Storytelling	0.57 (0.58/0.56)	0.48 (0.47/0.48)	0.57 (0.59/0.55)	0.64 (0.65/0.63)
Poetry	0.37 (0.42/0.33)	0.28 (0.34/0.24)	0.45 (0.47/0.43)	0.52 (0.57/0.48)
Theatre	0.72 (0.81/0.65)	0.69 (0.74/0.64)	0.81 (0.84/0.77)	0.85 (0.87/0.82)
Functionals	0.79 (0.75/0.83)	0.79 (0.77/0.82)	0.83 (0.81/0.84)	0.85 (0.83/0.87)
Interview	0.83 (0.87/0.85)	0.72 (0.70/0.75)	0.87 (0.84/0.90)	0.88 (0.86/0.91)
Others	0.14 (0.32/0.09)	0.13 (0.22/0.09)	0.23 (0.35/0.17)	0.23 (0.35/0.18)

Table 4. **F-measure (Precision/Recall computed in time percentage) for labeling each turn according to the six labels: Storytelling,Poetry,Theatre,Functionals,Interview,Others. Results are reported for prosodic and structural alone and in combination; the last column reports results whenever linear prediction features are also included.**

are reported in Table 4. In this case, the F-measures range in between 0.45 in case of poetry till 0.87 in case of interviews. As before, prosodic features outperform structural features on all the classes apart the functionals (presentations, conclusions, introduction and so on) where both features performs equally well (F-measure 0.79). As the stories labeled as 'Other' include segments without particular structure or speaking style their recognition rate is rather poor.

Figure 2 plots the confusion matrix in between the six different labels showing that while some classes are confidently detected (Theatre, Functionals, Interview), there is a large amount of confusion between Storytelling and Poetry labels where turns labeled as poetry in the reference are often assigned to storytelling labels. Next section addresses the problem of improving the discrimination between those two categories.

4.1 Prose versus Poetry

Both Storytelling and Poetry are characterized by narrative/expressive speaking styles, however the second is composed with more attention to rhythm and meters. The rhythm in poetry can appear by recurrent beat patterns of pauses across sentences or duration/stress of syllables similar to music [11]. Studies like [8] have shown how poetry patterns can be statistically modeled using a simple *linear regression* to predict the acoustic properties of segments (or a syllable) from the properties of preceding segments. In particular, authors showed in [8] that the Pearsons's r^2 coefficient, which quantifies the quality of the regression, is much higher in case of poetry then in case of prose. Furthermore in case of poetry, the Pearsons's r^2 coefficients are still high whenever the regression is estimated using up to the seventh preceding syllables/segments.

In order to investigate whether also folk poetry can be detected based on this principle, the average r^2 coefficients are included as features in the boosting. The coefficients extraction is based on the following procedure. At first pseudo-syllables are estimated as described in [14] segmenting the turns in a sequence of pseudo-syllable units $(u_1, u_2, ..., u_n)$. Also pauses are included in the sequence. Eight features are extracted, i.e, the unit durations, F0 statistics (maximum, minimum, mean, variance, slope), intensity and loudness, from each of them, thus producing eight feature sequences. For each unit and for each feature, a linear regression based on the seven preceding features is then estimated together with the r^2 coefficients as described in [8]. The r^2 coefficients are then averaged over the turn producing eight new features included in the booster. Table 4 (last column) reports as before F-measures (Precision/Recall) obtained in case of prosodic and structural features as well as after the inclusion of the Pearson's coefficients. It can be noticed that F-measure for Storytelling increases from 0.57 to 0.64 while F-measure for Poetry moves from 0.45 to 0.52. While the improvements are significant in both cases, F-measures are still lower than those obtained in case of remaining labels suggesting that in folk poetry, metric and rhythm regularities are not as marked as in case of data used in [8].

5. Discussion and Conclusion

Audio and speech processing methods have been already applied for automatically structuring, indexing and accessing spoken cultural heritage like interviews or witness-

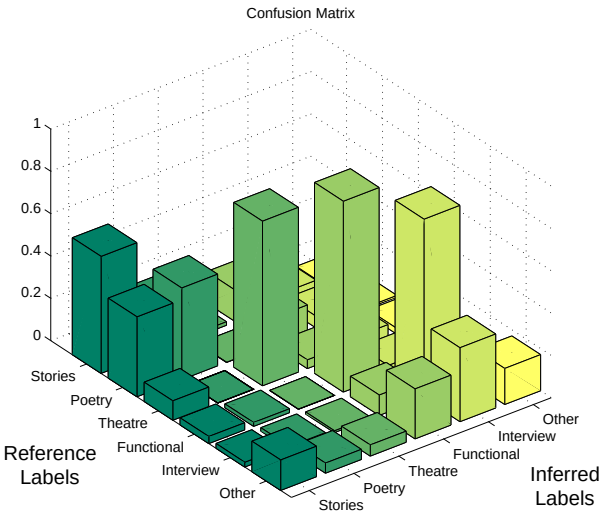

Figure 2. Confusion matrix between the six labels whenever boosting is trained/tested on prosodic and structural features.

ings [5, 12, 7]. This work further extends the use of those techniques for detecting and labeling folk literature in unstructured audio archives. The study is carried on 90 radio broadcast shows (35 hours of data), devoted to the preservation and disclosure of Swiss French dialects from various regions. The radio broadcast shows contain very different story types which includes interviews, tales, poetry, recitals, theatrical representations, discussions. In order to overcome the problem of dialect-dependent data sparseness, this work investigates the use of language-independent features based on the structure of the audio and on its acoustic/prosodic properties. Those features can be easily obtained by means of speaker diarization [2]. Results reveal that such a system can achieve an F-measure equal to 0.83 (Precision 0.87/ Recall 0.80) in retrieving folk literature versus more conventional broadcast data type. Prosodic features appear more effective and complementary to structural features. Furthermore, the paper investigates whether the same approach can be used to label speech segments into five large classes (Storytelling, Poetry, Theatre, Interviews, Functionals) showing F-measures ranging from 0.45 to 0.87. As last contribution, the paper investigates a novel set of features based on linear regression for modeling stylistic differences between prose and poetry. Including them in the classifiers increases the F-measures to the range 0.52 to 0.88.

In summary, through mean of language independent structural and prosodic features, it is possible to detect and label folk literature in large unstructured spoken audio archives with potential applications into accessing and disclosing those cultural heritage data.

References

[1] A. Adami and al. Qualcomm-icsi-ogi features for asr. In *Proceedings of ICSLP*, 2002.

[2] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[3] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind roles: Identifying speaker role in radio broadcasts. In *In Proc. AAAI 2000, The Seventeenth National Conference on Artificial Intelligence, Austin, Texas, USA*, pages 679–684, 2000.

[4] B. Bigot, J. Pinquier, I. Ferrane, and R. Andre-Obrecht. Looking for relevant features for speaker role recognition. In *INTERSPEECH*, 2010.

[5] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Wei-Jing Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, July 2004.

[6] J. Goldman and al. Accessing the spoken word. *Int. J. on Digital Libraries*, 5(4):287–298, 2005.

[7] J. H. L. Hansen and al. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing*, 13(5-1):712–730, 2005.

[8] G. Kochanski, A. Loukina, E. Keane, C. Shih, and B. Rosner. Long-range prosody prediction and rhythm. *Speech Prosody*, 2010.

[9] J. Llisterri. Speaking styles in speech research. *ESCA Workshop on Integrating Speech and Natural Language, Dublin, Ireland*, 2002.

[10] S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. *Proceedings of Eurospeech*, 2005.

[11] L. Nord, A. Kruckenberg, and F. Gunnar. Some timing studies of prose, poetry and music. *Speech Communication*, December 1990.

[12] R. J. F. Ordelman, W. F. L. Heeren, F. M. G. de Jong, M. A. H. Huijbregts, and D. Hiemstra. Towards affordable disclosure of spoken heritage archives. *Journal of Digital Information*, 10(6):17, December 2009.

[13] M. Ostendorf and al. Speech segmentation and its impact on spoken language technology. *IEEE Signal Processing Magazine*, 25(3), 2008.

[14] M. Piet. The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of Speech Prosody*, 2004.

[15] R. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 2000.

[16] E. Shriberg and al. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2), 2000.

[17] E. Shriberg and al. Svm modeling of snerf-grams for speaker recognition. *Proceedings of ICSLP*, 2004.

[18] M. Theune, K. Meijs, and D. Heylen. Generating expressive speech for storytelling applications. In *IEEE Transactions on Audio, Speech and Language Processing*, pages 1137–1144, 2006.