AUTOMATIC COMPOSITION OF BROADCAST NEWS SUMMARIES USING RANK CLASSIFIERS TRAINED WITH ACOUSTIC AND LEXICAL FEATURES

Taufiq Hasan[1], *Mohammed Abdelwahab*[2], *Srinivas Parthasarathy*[2], *Carlos Busso*[2] *and Yang Liu*[2]

[1]Research and Technology Center, Robert Bosch LLC, Palo Alto, CA, USA
[2]The University of Texas at Dallas, Richardson, TX, USA
taufiq.hasan@us.bosch.com, {busso,yang.liu}@utdallas.edu

ABSTRACT

Research on automatic speech summarization typically focuses on optimizing objective evaluation criteria, such as the ROUGE metric, which depend on word and phrase overlaps between automatic and manually generated summary documents. However, the actual quality of the speech summarizer largely depends on how the end-users perceive the audio output. This work focuses on the task of composing summarized audio streams with the aim of improving the quality and interest perceived by the end-user. First, using crowdsourced summary annotations on a broadcast news corpus, we train a rank-SVM classifier to learn the relative importance of each sentence in a news story. Acoustic, lexical and structural features are used for training. In addition, we investigate the perceived emotion level in each sentence to aid the summarizer in selecting interesting sentences, yielding an emotion-aware summarizer. Next, we propose several methods to combine these sentences to generate a compressed audio stream. Subjective evaluations are performed to evaluate the quality of the generated summaries on the following criterion: interest, abruptness, informativeness, attractiveness, and overall quality. The results indicate that users are most sensitive to the linguistic coherence and continuity of the audio stream.

Index Terms— Speech summarization, summary composition

1. INTRODUCTION

Automatic summarization of multimedia content has been an active area of research in the past decade. Summarization of spoken documents pose unique challenges since the information content is embedded in an audio signal, which raises the issue of errors in automatic speech recognition (ASR), speaker detection, sentence end-point detection, etc. [1, 2]. Generally, speech summarization aims at extracting the most informative and relevant sentences from the audio signal, and composing a concise version of the original spoken document [3–6]. The relevance of the selected sentences with respect to a human generated summary is evaluated using various objective quality metrics, such as the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [7].

Various supervised and unsupervised approaches have been proposed in the past for automatic speech summarization. Among unsupervised methods, maximal marginal relevance (MMR) [8], concept based integer linear programming (ILP) framework [9, 10], and graph based sub-modular selection approach [11] are noteworthy. In recent times, various supervised machine learning methods have been studied with encouraging results [3, 4, 12, 13]. Most of these techniques consider extractive summarization as a binary classification problem, i.e., a sentence either belongs to the summary or not. Various sentence-level features, including acoustic, prosodic, lexical and structural [3, 14], are used for this purpose. Alternative methods are also studied, where document summarization is seen as a sentence ranking problem [15–17]. In this approach, the relative importance among sentences within a document is learned from a human annotated summary corpora [18], and is later utilized to rank order the sentences of the document to be summarized. The process of constructing coherent summaries from the ranked sentences is still an open question. In contrast to text summarization, users are sensitive to acoustic/prosodic discontinuities in speech summaries.

In our previous work in [18], we proposed using rank-SVM techniques with crowdsourced summary annotations to rank sentences in a news story. In this work, we focus on composing an audio summary aiming to improve the perceived quality by a human listener. Unlike in [18], a full automatic scheme is presented for sentence segmentation and feature extraction. The proposed methods utilize the sentence ranking scores obtained from our rank-SVM classifier with automatic sentence segmentation and ASR transcripts [18]. We hypothesize that including emotional content in the speech summary will increase the interest of the listener and propose an emotion-aware speech summarizer. For this purpose, we obtain emotion labels from each sentence through crowdsourcing and generate a rank score based on emotion. These scores along with scores from the rank-classifier are then utilized to generate the summaries. We conduct subjective evaluations to compare the performance of the different summary composition methods and discuss the results.

2. CORPUS AND ANNOTATIONS

We utilized the RT-03 MDE Training Data Speech and annotations from Linguistic Data Consortium (LDC) [19]. This data-set consists of about 20 hours of Broadcast News along with transcriptions, sentence end-point labels and speaker information. We use a subset of 90 news stories as development data for the supervised speech summarization systems. In order to obtain human summary annotations, we utilize the crowdsourcing service Amazon Mechanical Turk (MTurk). The annotators are instructed to read the news story and select $10-15\%$ sentences that most effectively summarize the spoken document. Each story was annotated by 10 independent assessors. Further details can be found in [18].

For evaluation of the automatic summary generation methods, we utilize news stories independent from the development corpus. We collect online podcasts from National Public Radio (NPR), Wall street journal, and Cable News Network (CNN). Stories from the LDC corpus were also included. In total, 37 single story news segments were selected with average duration of $5-7$ minutes.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

Table 1. Experimental results comparing the baseline binary SVM system and the rank-SVM classifier using ROUGE-1, ROUGE-2 and ROUGE-L metrics. Lexical features are extracted using automatic speech recognition (ASR) engines.

		Metric and Compression Ratio (%)								
		ROUGE-1			ROUGE-2			ROUGE-L		
Classifier	Transcript	5%	10%	15%	5%	10%	15%	5%	10%	15%
Binary SVM	Manual	0.28258	0.45911	0.58506	0.18473	0.30051	0.39833	0.27147	0.44385	0.56900
	Google ASR	0.20116	0.43562	0.49554	0.11237	0.28191	0.30950	0.19090	0.42104	0.48014
	AT&T ASR	0.19597	0.42854	0.54463	0.10782	0.27752	0.35729	0.18648	0.41387	0.52815
Rank-SVM	Manual	0.35261	0.55974	0.68851	0.23974	0.39472	0.51296	0.33957	0.54450	0.67417
	Google ASR	0.31068	0.51539	0.64722	0.20727	0.35267	0.46334	0.29906	0.49943	0.63163
	AT&T ASR	0.34544	0.55404	0.68447	0.23389	0.39032	0.50779	0.33197	0.53841	0.66987

3. SENTENCE RANKING FOR SUMMARIZATION

The sentence ranking algorithm for summarization closely follows our previous work presented in [18]. We briefly describe the system here with additional results showing performance trade-off due to the use of ASR transcripts and contribution of different feature sets.

3.1. Acoustic, Lexical and Structural features

Various acoustic and prosodic features are extracted from each sentence as described in [20]. First, the short-term acoustic features, known as *Low level Descriptors (LLD)*, are extracted (e.g., fundamental frequency). Next, for each sentence in the corpus a global statistic of these features is calculated, yielding *high level features (HLF)*. The global statistic produces a single value for the entire sentence which is independent of the sentence length (e.g., mean of the fundamental frequency). The total feature set consists of 4368 dimensions which was reduced to 110 using a correlation based feature selection approach. Next, the following lexical features are extracted from each sentence: i) number of words, ii) number of Named Entities (NE), iii) number of stop-words, iv) sentiment polarity, v) TF-IDF (Term frequency - Inverse Document Frequency) vector, and v) bi-gram language model scores. Finally, the following structural features [21] are used in the summarization system: i) duration of the sentence, ii) duration of the sentence preceding the current sentence, iii) duration of the sentence following the current sentence, and iv) position of the current sentence within the story.

3.2. Classifiers

We utilize a *pair-wise* approach for ranking the sentences in the news story based on their relative importance [22–24]. Given the sentences in a story $\{s_1, s_2, ..., s_n\}$, the objective is to learn the preference relationship between pairs of instances and produce a rank score Λ_{rank} for each sentence. This score can be used to rank-order the sentences according to their relative importance. For comparison we train a binary-SVM classifier that decides whether a sentence is in the summary or not. The top 10% ranked sentences according to the annotator agreement counts are used as *summary* sentences and the rest as *non-summary* sentences.

Table 2. Experimental results comparing the performance of the different feature sets using the rank-SVM classifier with respect ROUGE metrics. Manual transcripts are used with $\%CR = 10$.

Feature set	ROUGE-1	ROUGE-2	ROUGE-L
Acoustic	0.51912	0.34858	0.50261
Lexical	0.52002	0.36904	0.50680
Structural	0.51642	0.32996	0.49939
All	0.55974	0.39472	0.54450

3.3. Objective Evaluation

We perform three set of experiments with the summary compression ratio (%CR) fixed at 5%, 10% and 15%, compared to the total number of sentences. The experimental protocol closely follows [18] with new results using ASR transcripts. Cloud ASR engines from Google [25] and AT&T [26] are utilized. For this particular corpus, the $\%$word error rate (WER) for the Google and AT&T ASR engines were found to be 19.33% and 23.18%, respectively. The ROUGE-1, ROUGE-2 and ROUGE-L metrics are used to evaluate the system summaries with respect to all 10 reference (annotated) summaries. The results are shown in Table 1, which compare the performances between manual transcriptions and ASR. The results include binary and rank-SVM to rank the sentences.

From the results in Table 1, consistent with our previous finding with manual transcripts [18], we again observe that the rank-SVM method provides superior results compared to the binary SVM classifier, even when ASR transcripts are used. As expected, a drop in ROUGE metrics is observed when ASR is used instead of manual transcripts. However, this drop does not seem to be significant, especially for the rank-SVM classifier.

We also compare the ROUGE metrics obtained from the summarizer when acoustic, lexical, and structural features are individually used. For completeness, we include the results for the rank-SVM trained with all the features. Here, we fix $\%CR = 10$ and only use the manual transcripts. From the results shown in Table 2, we observe that the lexical features perform the best, followed by acoustic and structural features. Also, using the full feature set provides better ROUGE performance than using each feature set individually.

3.4. Emotion based ranking

We hypothesize that emotional sentences are more interesting than non-emotional ones for speech summarization. To evaluate this hypothesis, we present an initial framework to modify the ranking of the sentences according to the emotional content of the sentences. While the emotional labels can be directly estimated from speech emotion classifiers [27], we decided to use labels assigned by humans during perceptual evaluations via crowdsourcing. Notice that this step is not fully automatic as the rest of the proposed system.

3.4.1. Emotion labels

We utilize Amazon Mechanical Turk to annotate each of the 37 news stories from the evaluation corpus. Each story is automatically segmented and the audio for each sentence was presented to the evaluators. Each sentence is emotionally evaluated by 5 evaluators, following the methodology used in Mariooryad *et al.* [28]. They were asked to rate the perceived emotional content of the sentence along two dimensions: arousal (excited vs. calm) and valence (positive vs. negative). These emotional attributes are annotated with pictorial representation to facilitate the understanding of these emotional dimensions (self-assessment manikins). We asked the evaluators: "How activated was the audio?" They were presented with five answer options ranging from very calm to very excited. For valence, we asked them: "How positive or negative is the sentence?" The answer options ranged from very negative to very positive. Finally, the answers were mapped into a numerical scale and the mean of the answers across evaluators for each sentence was calculated. The emotional dimension values range between -1 and 1 with neutral being around zero.

3.4.2. Emotional sentence scoring

Using the emotion labels, we generate a normalized rank score Λ_{emotion} from each sentence computed as:

$$\Lambda_{\text{emotion}} = \frac{\sqrt{Arousal^2 + Valence^2}}{\sqrt{2}} \tag{1}$$

Since emotion scores alone may not be reliable for selecting summary sentences, we fuse the rank-classifier scores Λ_{rank} with Λ_{emotion} as:

$$\Lambda_{\text{fusion}} = \alpha\Lambda_{\text{rank}} + (1-\alpha)\,\Lambda_{\text{emotion}} \tag{2}$$

We use $\alpha = 0.2$ for the fusion based on heuristics. We also use the extreme values of $\alpha = 0$ and 1.0, indicating only emotion and only classifier ranking, respectively.

4. AUTOMATIC COMPOSITION OF AUDIO SUMMARIES

In this section, we describe the automatic segmentation and summary generation methods. These methods utilize the sentence ranks obtained from the fused scores Λ_{fusion} and combines the extracted sentences in order to compose a concise, linguistically meaningful, naturally flowing and interesting summary.

4.1. Automatic segmentation

The segmentation process has three main steps. First, we use the LIUM speaker diarization system [29] to segment the recordings into speaker's turns. Second, we use a two class supervised clustering based on the expectation maximization (EM) to remove music segments. Third, we split the speakers' turns into sentences. For this purpose, voice activity detection [30, 31] is performed as a parallel step on the entire audio stream. We utilize the silence regions in the broadcast news audio stream to segment the speakers' turns into sentences. Notice that only some of the silences' breaks correspond to sentence boundaries. Therefore, a decision tree classifier was built based on several speech and structural features to identify sentence boundaries following the work proposed by Shriberg *et al.* [32].

4.2. Summary composition methods

This section describes the proposed methods for creating news summaries based on the ranking of the sentences. Trying to include separate sentences inadvertently introduces "cuts" or "discontinuities" which cannot be resolved by simple means. We also asked an expert audio engineer to manually summarize the news stories in order to use them as gold-standards for summary composition.

4.2.1. Anchor's summary

In this method, we utilize the first continuous segment of the news audio where the anchor usually provides an overview of the news story. We concatenate each sentence from the beginning of the story until one of the following conditions is met: i) a significant drop of rank is observed, ii) the user defined length limit is reached, iii) the speaker has changed. In this method, there is no loss in continuity of speech or the context but the most informative sentences, possibly buried in the middle of the story, may not be included.

4.2.2. Trailer-like summary

In this method, the first sentence is always selected, which is then followed by chronologically combining high rank sentences until the duration limit is reached. In order to reduce discontinuity, sentences before and after the selected sentences are also included. Obviously, all the sentences selected here will not be consecutive (in the original story). Therefore, coherence and continuity may be compromised. These summaries provide a trailer-like feel to the news, with an introductory phrase followed by important/interesting portions within the audio.

4.2.3. Hot-spot summary

In this form of summary, the region surrounding the highest ranked sentence is identified. The summary begins from the speaker's segment that includes the most important sentence in order to minimize incoherence and continues until the duration limit is reached. This method is similar to the trailer method but provides more continuity and improved flow. However, without the introductory sentence, the summary may seem to lack proper context.

4.2.4. Order-based summary

In this method the summary is formed by simply combining the sentences of highest ranking until the limit is reached. This method provides the most important sentences. However, the summary may be less coherent with abrupt transitions.

4.2.5. Human composed summary

A highly skilled audio engineer (trained on sound design and music composition) carefully listened to the audio and crafted these summaries aiming to optimize coherence and listener satisfaction. The audio engineer utilizes full or partial sentences and merged them in any order that sounds appropriate and meaningful. These are considered gold-standard for audio summaries.

5. SUBJECTIVE EVALUATIONS

This section describes the subjective evaluation experiments on the automatically composed summaries. We utilize the 37 stories from the evaluation corpus. Since manual transcripts are not available for

Table 3. Subjective evaluation criteria

Name	**Description**	**Numeric scale**	**Subjective scale**
Interest	How interesting is the summary?	$[-2, 2]$	[Extremely uninteresting , $\cdots$, Very Interesting]
Abruptness	How often did you notice unusual discontinuities in the summary?	$[-2, 2]$	[Very frequently, $\cdots$, None]
Information	Does the summary provide adequate information about the story?	$[-2, 2]$	[Extremely inadequate, $\cdots$, Very adequate]
Attractiveness	How likely are you to listen to entire news story after hearing the summary?	$[-2, 2]$	[Extremely unlikely, $\cdots$, Very likely]
Quality	How is the overall quality of the audio?	$[-2, 2]$	[Very Bad, $\cdots$, Very Good]

Table 4. Objective scores of summary generation methods using ROUGE-1, ROUGE-2 and ROUGE-L metrics.

Metric	ROUGE-1	ROUGE-2	ROUGE-L
Anchor	0.70405	0.53151	0.69384
Hotspot	0.52227	0.27462	0.48587
Order-based $\alpha = 0.0$	0.3564	0.08673	0.32697
Order-based $\alpha = 0.2$	0.40873	0.15622	0.37129
Order-based $\alpha = 1.0$	0.51195	0.26211	0.47213
Trailer-like $\alpha = 0.0$	0.63358	0.43707	0.60678
Trailer-like $\alpha = 0.2$	0.63793	0.43578	0.61311
Trailer-like $\alpha = 1.0$	0.62464	0.43873	0.59256

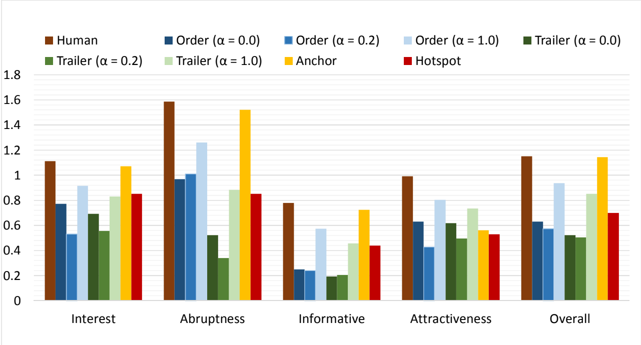

Fig. 1. Average subjective evaluation scores for the summary generation methods. Results are shown in the five user criteria: i) interest, ii) abruptness, iii) informative, iv) attractiveness, and v) overall.

these news segments, ASR was used for speech-to-text conversion. The segmentation was done in a fully automatic fashion as discussed in Sec. 4.1. The rank-SVM classifier trained on the LDC data was used to rank the sentences. Summaries were prepared from all 37 stories using the methods described in the previous section.

Perceived quality of audio and user's satisfaction differs among users depending on their personalities and interests [33]. In order to mitigate such variations, we use Amazon Mechanical Turk to recruit a large number of participants. A total of 10 evaluators listened to the automatic summaries while a manual transcript of the full news story is provided to them for reference. We use five different evaluation criteria to measure the quality of each news summary. These are described in Table 3 along with the numeric and subjective scales used during the evaluation. To ensure the quality of the results, the questions appeared to the evaluators only after they finished listening to the news summaries.

5.1. Results

The averaged subjective evaluation results are shown in Fig. 1 across the five criteria (Table 3). As expected, the human expert generated summary provides the highest scores in all criteria. The anchor's summary provides the second best scores, except on atractiveness. Trailer ($\alpha = 1.0$) method shows slight superiority in the attractiveness category compared to anchor's summary, meaning that the user is more likely to listen to the entire news story after listening to the summary. This is an intended effect of the trailer approach, where important sentences included from the middle of the story creates a sense of curiosity in the listeners. Fig. 1 further illustrates that using emotion does not necessarily provide any benefit to the perceived quality of the summary. While for news summary we observe this result, other domains may benefit from leveraging emotional content. Overall, anchor's summary seems to be the best full automatic method judging by the subjective evaluation.

We use the human generated summary as ground truth, and compute the ROUGE metrics for the summary generation methods (objective metric for the results). Table 4 presents the results. We observe that the Anchor's summary correlates the most with human summary. This makes sense given that in news domain, most of the relevant information is contained in the beginning of the segment. However contradictory to the subjective results, the trailer-like summary outperforms the order-based summary. This means that the information contained in the trailer is closer to the human summary than that of the order-based summary. According to the results, utilizing emotion did not provide any significant advantage in summary composition. We believe this was due to the lack of sufficient emotional variability in the news stories.

6. CONCLUSIONS

In this work, we have proposed an automated summary generation system that utilized rank classifiers for selecting summary sentences. We used a large set of acoustic features, along with conventional lexical and structural features. Crowdsourced summary annotations have been used to train the rank classifiers. We have also proposed various fully automatic summary generation methods and studied the effect of deliberately including emotional content in the speech summary. Experiments have been performed to identify the advantages and drawbacks of the proposed methods along with different subjective and objective evaluation criteria. The evaluation provides promising results to generate automatic speech summary, overcoming the challenges in combining relevant sentences. While including emotion information in ranking the sentences did not improve the results on news, a future work is to evaluate this framework in other domains where the recordings are emotionally colored (e.g., movie review, political debates).

7. REFERENCES

[1] Inderjeet Mani and Mark T Maybury, *Advances in automatic text summarization*, vol. 293, MIT Press, 1999.

[2] Ani Nenkova, Sameer Maskey, and Yang Liu, "Automatic summarization," in *Proc. ACL: Tutorial Abstracts*. ACL, 2011, p. 3.

[3] Sameer Maskey and Julia Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization.," in *Proc. Interspeech*, 2005, pp. 621–624.

[4] Berlin Chen, Shih-Hsiang Lin, Yu-Mei Chang, and Jia-Wen Liu, "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing & Management*, vol. 49, no. 1, pp. 1–12, 2013.

[5] Fei Liu and Yang Liu, "Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no. 7, pp. 1469–1480, 2013.

[6] Shasha Xie and Yang Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, vol. 24, no. 3, pp. 495–514, 2010.

[7] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.

[8] Klaus Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[9] Daniel Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proc. IEEE ICASSP*. IEEE, 2009, pp. 4769–4772.

[10] Shasha Xie, Dilek Hakkani-Tur, Benoit Favre, and Yang Liu, "Integrating prosodic features in extractive meeting summarization," in *Proc. IEEE ASRU*. IEEE, 2009, pp. 387–391.

[11] Shih-Hsiang Lin and Berlin Chen, "Improved speech summarization with multiple-hypothesis representations and kullback-leibler divergence measures.," in *Proc. Interspeech*, 2009, pp. 1847–1850.

[12] Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey, "From text to speech summarization," in *Proc. IEEE ICASSP*, 2005, pp. 997–1000.

[13] Shih-Hsiang Lin, Berlin Chen, and Hsin-Min Wang, "A comparative study of probabilistic ranking models for chinese spoken document summarization," *ACM Trans. on Asian Language Information Process.*, vol. 8, no. 1, pp. 3, 2009.

[14] Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore, "Evaluating automatic summaries of meeting recordings," in *In Proc. ACL*, 2005.

[15] Shih-Hsiang Lin, Yu-Mei Chang, Jia-Wen Liu, and Berlin Chen, "Leveraging evaluation metric-related training criteria for speech summarization," in *Proc. IEEE ICASSP*. IEEE, 2010, pp. 5314–5317.

[16] Yueng-Tien Lo, Shih-Hsiang Lin, and Berlin Chen, "Constructing effective ranking models for speech summarization," in *Proc. IEEE ICASSP*. IEEE, 2012, pp. 5053–5056.

[17] Yi-Ting Chen, Berlin Chen, and Hsin-Min Wang, "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 1, pp. 95–106, 2009.

[18] Srinivas Parthasarathy and Taufiq Hasan, "Automatic broadcast news summarization via rank classifiers and crowdsourced annotation," in *Proc. IEEE ICASSP*. IEEE, 2015.

[19] Strassel Stephanie, Christopher Walker, and Haejoong Lee, "RT-03 MDE Training Data Speech LDC2004S08," [Online] `https://catalog.ldc.upenn.edu/LDC2004S08`, 2004.

[20] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski, "The interspeech 2011 speaker state challenge.," in *Proc. Interspeech*. ISCA, 2011, pp. 3201–3204.

[21] Sameer Maskey and Julia Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proc. Interspeech*, 2003.

[22] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM SIGKDD*. ACM, 2002, pp. 133–142.

[23] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *Proc. ACM ICML*. ACM, 2005, pp. 89–96.

[24] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li, "Learning to rank: from pairwise approach to listwise approach," in *Proc. ACM ICML*. ACM, 2007, pp. 129–136.

[25] "Google Speech API v2," `https://github.com/gillesdemey/google-speech-v2`, Accessed: 03-25-2015.

[26] "AT&T Speech API," `http://developer.att.com/apis/speech`, Accessed: 03-25-2015.

[27] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[28] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[29] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," Tech. Rep., Idiap, 2013.

[30] Pavel Matejka, Petr Schwarz, Jan Cernockỳ, and Pavel Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. Interspeech*, 2005, pp. 2237–2240.

[31] Igor Szöke, Petr Schwarz, Pavel Matejka, Lukas Burget, Martin Karafiát, Michal Fapso, and Jan Cernockỳ, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. Interspeech*, 2005, pp. 633–636.

[32] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech communication*, vol. 32, no. 1, pp. 127–154, 2000.

[33] Kari Kallinen and Niklas Ravaja, "Emotion-related effects of speech rate and rising vs. falling background music melody during audio news: The moderating influence of personality," *Personality and individual differences*, vol. 37, no. 2, pp. 275–288, 2004.