Prosody-Based Unsupervised Speech Summarization with Two-Layer Mutually Reinforced Random Walk

Sujay Kumar Jauhar, Yun-Nung Chen, and Florian Metze
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213-3891, USA
{sjauhar,yvchen,fmetze}@cs.cmu.edu

Abstract

This paper presents a graph-based model that integrates prosodic features into an unsupervised speech summarization framework without any lexical information. In particular it builds on previous work using mutually reinforced random walks, in which a two-layer graph structure is used to select the most salient utterances of a conversation. The model consists of one layer of utterance nodes and another layer of prosody nodes. The random walk algorithm propagates scores between layers to use shared information for selecting utterance nodes with highest scores as summaries. A comparative evaluation of our prosody-based model against several baselines on a corpus of academic multi-party meetings reveals that it performs competitively on very short summaries, and better on longer summaries according to ROUGE scores as well as the average relevance of selected utterances.

1 Introduction

Automatic extractive speech summarization (Hori and Furui, 2001) has garnered considerable interest in the natural language processing research community for its immediate application in making large volumes of multimedia documents more accessible. Several variants of speech summarization have been studied in a range of target domains, including news (Hori et al., 2002; Maskey and Hirschberg, 2003), lectures (Glass et al., 2007; Chen et al., 2011) and multi-party meetings (Banerjee and Rudnicky, 2008; Liu and Liu, 2010; Chen and Metze, 2012b).

Research in speech summarization – unlike its text-based counterpart – carries intrinsic difficulties, which draw their origins from the noisy nature of the data under consideration: imperfect ASR transcripts due to recognition errors, lack of proper segmentation, etc. However, it also offers some advantages by making it possible to leverage extra-textual information such as emotion and other speaker states through an incorporation of prosodic knowledge into the summarization model.

A study by Maskey and Hirschberg (2005) on the relevance of various levels of linguistic knowledge (including lexical, prosodic and discourse structure) showed that enhancing a summarizer with prosodic information leads to more accurate and informed results.

In this work we extend the model proposed by Chen and Metze (2012c), where a random walk is performed on a lexico-topical graph structure to yield summaries. They exploited intra- and inter-speaker relationships through partial topic sharing for judging the importance of utterances in the context of multi-party meetings. This paper, on the other hand, enriches the underlying graph structure with prosodic information, rather than lexico-topical knowledge, to model speaker states and emotions.

Also different from Maskey and Hirschberg (2005), we model the multimedia document structure as a graph, which allows for flexibility as well as expressive power in representation. This graph structure provides the easy incorporation of targeted features into the model as well as in-depth analyses of individual feature contributions towards representing speaker information.

To the best of our knowledge this paper presents the first attempt at performing speech summarization using no lexical information in a completely unsupervised setting. Maskey and Hirschberg (2006) use an HMM to perform summarization by relying solely on prosodic features. However, their model – unlike ours – is supervised. The only requirement of the model in this paper is a preprocessing step that segments the audio into "ut-

International Joint Conference on Natural Language Processing, pages 648–654,
Nagoya, Japan, 14-18 October 2013.

terances".

While utterance segmentation may be a non-trivial problem, the possibility of an unsupervised speech summarization model that relies solely on acoustic input is advantageous. Importantly, it does not rely on any training data and circumvents the primary difficulties that plague most speech summarization techniques — namely the noise introduced into the system by imperfect speech recognition.

We evaluate our model on a dataset consisting of multi-party academic meetings (Chen and Metze, 2012b; Banerjee and Rudnicky, 2008). We perform evaluation using the ROUGE metric for automatic summarization, which counts n-gram overlap between reference and candidate summaries. We also run a post-hoc analysis, which measures the average relevance score of utterances in a candidate summary.

Evaluation results indicate that our model outperforms a number of baselines across varying experimental settings in all but the shortest summaries. We hence claim that our model is a robust, flexible, and effective framework for unsupervised speech summarization.

The rest of the paper is organized as follows. Section 2 describes the prosodic features encoded in the model and how they are extracted. Section 3 presents the construction of the two-layer graph and mutually reinforced random walk for propagating information through the graph. Section 4 shows experimental results of applying the proposed model to the dataset of academic meetings and discusses the effects of prosody on summarization. Section 5 concludes.

2 Prosodic Feature Extraction

As previously stated, the only pre-requisite of the model proposed in this paper is a segmentation of the input document into chunks that are dictated by some meaningful notion of utterances. Once the audio has been segmented utterance-wise, the rest of the pipeline is effectively agnostic to all but its acoustic properties.

Given a set of pre-segmented audio files, we extract the following prosodic features from them using PRAAT scripts (Huang et al., 2006).

- Number of syllables and number of pauses.
- Duration time, – which is the speaking time including pauses – and the phonation time, – which is the speaking time excluding pauses.
- Speaking rate and articulation rate, which are the number of syllables divided by the duration time and phonation time, respectively.
- The average, maximal and minimal fundamental frequencies measured in Hz (which objectify the perceptive notion of pitch).
- The energy measured in Pa^2/sec and the intensity measured in dB.

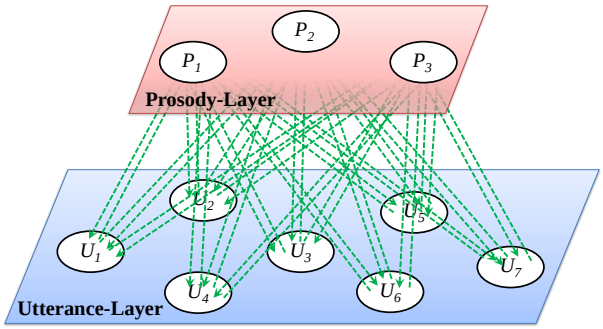

Figure 1: *A simplified example of the two-layer graph considered, where a type of prosody P_i is represented as a node in prosody-layer and an utterance U_j is represented as a node in utterance-layer of the two-layer graph.*

The inclusion of the features above into the model was motivated by their possible contribution to the notion of "important utterances" in a dialogue. For example, intuitively, pitch is a vocal channel for emotions, such as anger, or embarrassment. It may thus contribute, via the emotional investment of the speaker to the importance of her utterances. Similarly, the variation of energy over an utterance results in its perceived loudness, thus possibly permitting the inference of emphasis or stress to particular utterances by speakers. Again, speech rate often acts as a latent channel for communication of information, where excitement or emphasis is implicitly conveyed by a speaker.

3 Two-Layer Mutually Reinforced Random Walk

In this section we describe our method for modelling speech data as a two-layered interconnected graph structure and run the mutually reinforced random-walk algorithm for summarization.

Given an input speech document that is suitably segmented into utterance chunks, we construct a linked two-layer graph G containing an utterance set V_U and a prosody set V_P. Each node of the graph $U_i \in V_U$ corresponds to a single utterance as obtained from the pre-processing "chunking" step. Every node $P_i \in V_P$ illustrates a single prosodic features incorporated into the model.

Figure 1 shows a simplified example of such a two-layered graph. $G = \langle V_U, V_P, E_{UP}, E_{PU} \rangle$, where $V_U = \{U_i\}$, $V_P = \{P_i\}$, $E_{UP} = \{e_{ij} \mid U_i \in V_U, P_j \in V_P\}$, and $E_{PU} = \{e_{ij} \mid P_i \in V_P, U_j \in V_U\}$. Here, E_{UP} and E_{PU} represent the sets of directional edges between utterances and prosodic nodes with different directions (Cai and Li, 2012).

Based on these sets of directional edges we further define $L_{UP} = [w_{i,j}]_{|V_U| \times |V_P|}$ and $L_{PU} = [w_{j,i}]_{|V_P| \times |V_U|}$. The matrices L_{UP} and L_{PU} effectively encode the directional relationship between utterances and prosodic features. More concretely, for example, the entry $w_{i,j}$ of L_{UP} is the value of the prosodic feature P_j extracted from the utterance U_i. Row-normalization is performed on L_{UP} and L_{PU} (Shi and Malik, 2000). It may be noted that, as a consequence, L_{UP} is different from L_{PU}^T.

Traditional random walk only operates on a single layer of the graph structure and integrates the initial similarity scores with the scores propagated from other utterance nodes (Chen et al., 2011; Chen and Metze, 2012a; Hsu et al., 2007). The approach adopted in this paper, however, considers prosodic information by propagating information between layers based on external mutual reinforcement (Chen and Metze, 2012c).

Effectively the working of the algorithm stems from two interrelated intuitions. On the one hand, utterances that evidence more pronounced signs of important prosodic features should themselves be judged as more salient. On the other hand, prosodic features in salient utterances that are recorded with higher values should themselves be deemed as more important.

The advantage of the algorithm is that it is entirely unsupervised and allows for the integration of knowledge-rich target specific features. The mathematical formulation of the algorithm is presented as follows.

Given some initial scores $F_U^{(0)}$ and $F_P^{(0)}$ for utterance and prosody nodes respectively, the update rule is given by:

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases} \quad (1)$$

Here $F_U^{(t)}$ and $F_P^{(t)}$ integrate the initial importance associated with their respective nodes with the score obtained by between-layer propagation at a given iteration t.

Hence, the scores in each layer are mutually updated by the scores from the other layer, iteratively. In particular, utterances that exhibit more pronounced signs of important prosodic feature are progressively scored higher. At the same time, prosodic features that appear with higher values in salient utterances become progressively more important.

For the utterance set, the update rule increments the importance of nodes with the combination $L_{UP}F_P^{(t)}$. This latter term can be considered as the score from linked nodes in the prosody set, weighted by prosodic feature values. Finally, an α value encodes the trade-off between initial utterance weight and information sharing via propagation. The algorithm converges satisfying (2).

$$\begin{cases} F_U^{*} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{*} \\ F_P^{*} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{*} \end{cases} \quad (2)$$

Additionally F_U^* has an analytical solution which is given by:

$$\begin{aligned} F_U^* &= (1-\alpha)F_U^{(0)} \\ &+ \alpha \cdot L_{UP}\left((1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^*\right) \\ &= (1-\alpha)F_U^{(0)} + \alpha(1-\alpha)L_{UP}F_P^{(0)} \\ &+ \alpha^2 L_{UP}L_{PU}F_U^* \\ &= \left((1-\alpha)F_U^{(0)}e^T + \alpha(1-\alpha)L_{UP}F_P^{(0)}e^T \right. \\ &+ \left. \alpha^2 L_{UP}L_{PU}\right)F_U^* \\ &= MF_U^*, \end{aligned} \quad (3)$$

where $e = [1, 1, ..., 1]^T$. The closed-form solution F_U^* of (3) is the dominant eigenvector of M (Langville and Meyer, 2005).

It may be noted that for the practical implementation of the algorithm, we set the initial scores of utterance nodes $F_U^{(0)}$ and prosodic nodes $F_P^{(0)}$ to have equal importance. Also we empirically set $\alpha = 0.9$ for all our experiments because several studies have shown that $(1-0.9)$ is a proper damping factor (Hsu et al., 2007; Brin and Page, 1998).

4 Experiments

4.1 Pre-processing – Time Alignment

We have previously stressed that while our model is independent from the lexical representation of an audio document, it does rely on a preprocessing step that chunks the document into individual utterances. It is noted that this may not be a trivial task.

Speaker diarization (Tranter and Reynolds, 2006) and utterance segmentation (Christensen et al., 2005; Geertzen et al., 2007) are open areas of research in the NLP community. Systems developed for these purposes may be used to produce the initial chunking required by our model. In this paper, however, we do not explore these methods and instead rely on segmentation obtained from manually produced textual transcripts. This is to study the efficacy of our model in isolation.

A second reason for using textual transcripts is the presentation of experimental evaluation. This form of data allows for tangible results that are obtained through evaluation metrics such as ROUGE, which rely on measuring n-gram overlap between reference and candidate summaries. Furthermore, the resulting textual surface form and summaries are more "semantically" interpretable as well.

To associate prosodic information with the textual realization of each utterance in a manual transcript, a preprocessing step requires time alignments between the audio and the corresponding text of each utterance. Note that this step is unnecessary in the case when manual transcripts are not present, and utterance chunking is obtained from some other, automatic means. The time alignment is then implicitly obtained in the process of utterance segmentation.

To accomplish the alignment in our experimental framework, a speech recognizer is first used to produce an ASR output of the audio document. A by-product of this step is that each recognized token contains an inherent time signature. Using Viterbi alignment between the ASR output and manual transcription the time signatures from the audio is projected onto each manually transcribed utterance.

We experimented with Viterbi alignment at a number of different levels of granularity including token level, character level, and phoneme level (via conversion of text to phonetic representation using the CMU pronunciation dictionary (Weide, 1998)). The latter was empirically found to produce the most fine-grained and precise alignments, and was consequently used in all our experiments.

4.2 Corpus

The dataset used in our evaluation is the same one previously employed by Chen and Metze (2012b). It consists of 10 meetings held between April and June 2006, with largely overlapping participants and topics of discussion. There were a total of 6 unique participants, with each meeting involving between 2 and 4 individual speakers. SmartNotes (Banerjee and Rudnicky, 2008) was use to record both the audio and the notes for each meeting.

The average duration of a meeting in the dataset was approximately 28 minutes, and the total number of utterances was 7123. We only use the manual transcripts of the meetings to actually evaluate our model, although ASR transcripts were used for time alignment.

The reference summaries are produced by selecting the set of the most "noteworthy" utterances. Two annotators manually labelled the degree of "noteworthiness" (on a relevance scale of 1 to 3) for each utterance. We extract all the utterances with a "noteworthiness" level of 3 to form the reference summary of each meeting.

4.3 Baselines

Several baselines were used for comparison against our model and are described below.

1. **Longest:** The first baseline simply selects the longest utterances to form a summary of a document (where the length of the extracted summary is based on the desired ratio). We define the length of utterances by the number of tokens they contain.

2. **Begin:** A second variant of this baseline selects the utterances that appear in the beginning of the document.

3. **LTE:** The third baseline is a summary produced by using Latent Topic Entropy (LTE) (Kong and Lee, 2011). This measure essentially estimates the "focus" of an utterance. Hence, theoretically, a lower topic entropy relates to a more topically informative utterance, which in turn translates into a noteworthy utterance to include in a summary.

4. **TF-IDF** The final baseline uses basic TF-IDF to measure the importance of utterances, by

F-measure		ROUGE-1			ROUGE-L		
		10%	20%	30%	10%	20%	30%
Baseline	Longest	34.05	52.48	61.11	33.66	52.10	60.77
	Begin	**35.45**	54.42	64.63	**35.28**	54.18	64.37
	LTE	35.16	54.67	64.97	35.03	54.54	64.76
	TFIDF	32.01	51.33	63.11	31.89	51.08	62.84
This Paper		35.33	**55.17**	**65.60**	35.09	**54.90**	**65.36**

Table 1: ROUGE scores (%) on multi-party meeting dataset

taking the averaged TF-IDF score over each of its individual words.

It may be noted that the topic distribution of words as well as their IDF scores were obtained by computing statistics over all ten meetings in our experimental dataset.

4.4 Evaluation Metrics

Our automated evaluation utilizes the standard DUC (Document Understanding Conference) evaluation metric, ROUGE (Lin, 2004), which measures recall over various n-gram statistics between a system-generated summary and a set of summaries produced by humans. F-measures for ROUGE-1 (unigram) and ROUGE-L (longest common subsequence) can be evaluated in exactly the same way.

We also use a post-hoc evaluation metric to measure the average "importance" of utterances in a summary. This metric associates a relevance score to a summary by taking the averaged noteworthiness score of each utterance, as obtained from human annotators.

4.5 Results and Discussion

We ran each of the baseline summarizers as well as the system proposed in this paper to produce 10%, 20% and 30% summaries of each of the meetings in the dataset. The percentage of a summary was determined by selecting the top k utterances (as determined by a given system) until the desired ratio between the number of tokens in the summary to the total length of its corresponding meeting was met.

Evaluation results on the ROUGE metric are presented in Table 1. They reveal that the performance of our prosody-based model is competitive with the other baselines on the shortest 10% summaries. In fact it ranks second, only scoring lower than the baseline that considers the beginning of a document as a summary. Additionally, on the longer 20% and 30% summaries, the system outperforms all the baselines.

We believe that in the case of very short summaries, the nature of the data under consideration biases the evaluation of the "begin" baseline. This is because the meetings generally commence with a presentation of an agenda which contains key terms that are likely to be discussed during the course of the rest of the session. In this scenario a metric such as ROUGE – which effectively measures n-gram overlap – would reward the "begin" summary for including key terms that appear several times in the gold standard summaries.

However for longer summaries, where lexical variation is more pronounced, prosodic information provides a robust source of intelligence to select noteworthy utterances. In fact we are surprised that it outperforms the lexically derived LTE and TF-IDF baselines in all evaluation configurations.

Overall, these results seem to suggest that our model is able to capture latent speaker information and incorporate it effectively into the process of extractive summarization.

We further test this conclusion by conducting a post-hoc analysis, where we examine the average "importance" of utterances in the summary produced by a particular system. More specifically, we measure the average relevance score – ranging on a scale of 1 to 3 – of the utterances, where the score of each utterance is derived from its noteworthiness level as judged by human annotators (Banerjee and Rudnicky, 2008). The results of this analysis are presented in Table 2.

While the "begin" baseline is able to produce summaries with the highest relevance score for the shortest 10% summaries, our model outperforms all other systems on the longer 20% and 30% summaries. Moreover, it is competitive with the "begin" baseline even on the shortest summaries and scores higher than the other baselines. These re-

Avg. Relevance		10%	20%	30%
Baseline	Longest	2.299	2.272	2.283
	Begin	**2.464**	2.402	2.398
	LTE	2.334	2.369	2.367
	TFIDF	2.355	2.363	2.375
This paper		2.454	**2.422**	**2.411**

Table 2: Avg. relevance scores on multi-party meeting dataset

sults align with the findings in Table 1.

As an auxiliary analysis we also extract the converged scores of prosody nodes and rank them in order to analyze their effectiveness. The ranking reveals that the number of pauses in an utterance, its minimum and average pitch, and its intensity tend to be the most predictive features. In the context of academic meetings the number of pauses may be indicative of the time a speaker takes to formulate and articulate his/her thoughts. Thus more pauses may indicate utterances that have been more carefully crafted and therefore include more relevant content. Pitch and intensity are generally good measures of important information, because speakers tend to use them to express emotion. This fact has previously been successfully leveraged for key term extraction (Chen et al., 2010).

Conversely the duration time of the utterance, the number of syllables, and the energy are the least predictive features. With the exception of energy, the other two features can be considered as a surrogate measure for the length of utterances. This parallels what the "longest" utterance baseline performs lexically. The finding corresponds to the results from Tables 1 and 2, which show that this baseline does not produce particularly relevant summaries.

5 Conclusion

Our paper proposes a novel approach to integrating speaker-state information, through the incorporation of prosodic knowledge into an unsupervised model for extractive speech summarization. We have also shown the first attempt at performing unsupervised speech summarization without using lexical information.

We have presented experiments on a dataset of academic meetings involving spoken interactions between multiple parties. Evaluation results indicate that our model extracts relevant utterances as summaries, both from the perspective of automatic evaluation metrics such as ROUGE as well as a post-hoc metric that measured the average relevance score of utterances within summaries. In addition our model compared favorably with a number of heuristic and lexically derived baselines outperforming them in all but one scenario. This substantiates its claim to a robust and viable method for completely unsupervised speech summarization.

References

Satanjeev Banerjee and Alexander I. Rudnicky. 2008. An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. In *Proceedings of The 2nd IEEE Workshop on Spoken Language Tachnology (SLT)*, pages 177–180. IEEE.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Xiaoyan Cai and Wenjie Li. 2012. Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization. *IEEE Transactions on Acoustics, Speech and Language Processing*, 20:1597–1607.

Yun-Nung Chen and Florian Metze. 2012a. Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in random walk for improved multi-party meeting summarization. In *Proceedings of The 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Yun-Nung Chen and Florian Metze. 2012b. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In *Proceedings of The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 377–381, Montréal, Canada, June. Association for Computational Linguistics.

Yun-Nung Chen and Florian Metze. 2012c. Two-layer mutually reinforced random walk for improved multi-party meeting summarization. In *Proceedings of The 4th IEEE Workshop on Spoken Language Tachnology (SLT)*.

Yun-Nung Chen, Yu Huang, Sheng-Yi Kong, and Lin-Shan Lee. 2010. Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 265–270. IEEE.

Yun-Nung Chen, Yu Huang, Ching-Feng Yeh, and Lin-Shan Lee. 2011. Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms. In *Proceedings of The 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Heidi Christensen, BalaKrishna Kolluru, Yoshihiko Gotoh, and Steve Renals. 2005. Maximum entropy segmentation of broadcast news. *IEEE Signal Processing Society Press.*

Jeroen Geertzen, Volha Petukhova, and Harry Bunt. 2007. A multidimensional approach to utterance segmentation and dialogue act classification. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp*, pages 140–149.

James R Glass, Timothy J Hazen, D Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay. 2007. Recent progress in the mit spoken lecture processing project. In *Proceedings of The 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2553–2556.

Chiori Hori and Sadaoki Furui. 2001. Advances in automatic speech summarization. *RDM*, 80:100.

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex" Waibel. 2002. Automatic speech summarization applied to english broadcast news speech. In *Proceedings of ICASSP*, volume 1, pages 9–12.

Winston H Hsu, Lyndon S Kennedy, and Shih-Fu Chang. 2007. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th international conference on Multimedia*, pages 971–980. ACM.

Zhongqiang Huang, Lei Chen, and Mary Harper. 2006. An open source prosodic feature extraction tool. In *Proceedings of the Language Resources and Evaluation Conference*.

Sheng-Yi Kong and Lin-Shan Lee. 2011. Semantic analysis and organization of spoken documents based on parameters derived from latent topics. *IEEE Trans. on Audio, Speech and Language Processing*.

Amy N Langville and Carl D Meyer. 2005. A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1):135–161.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Fei Liu and Yang Liu. 2010. Using spoken utterance compression for meeting summarization: A pilot study. In *Proceedings of The 3rd IEEE Workshop on Spoken Language Tachnology (SLT)*.

Sameer Raj Maskey and Julia Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *Proceedings of Eurospeech*.

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proceedings of InterSpeech*.

Sameer Maskey and Julia Hirschberg. 2006. Summarizing speech without text using hidden markov models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 89–92. Association for Computational Linguistics.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565.

RL Weide. 1998. The cmu pronunciation dictionary, release 0.6.