Incorporating Lexical and Prosodic Information at Different Levels for Meeting Summarization

Catherine Lai, Steve Renals

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, UK
clai@inf.ed.ac.uk, srenals@ed.ac.uk

Abstract

This paper investigates how prosodic features can be used to augment lexical features for meeting summarization. Automatic detection of summary-worthy content using non-lexical features, like prosody, has generally focused on features calculated over dialogue acts. However, a salient role of prosody is to distinguish important words within utterances. To examine whether including more fine grained prosodic information can help extractive summarization, we perform experiments incorporating lexical and prosodic features at different levels. For ICSI and AMI meeting corpora, we find that combining prosodic and lexical features at a lower level has better AUROC performance than adding in prosodic features derived over dialogue acts. ROUGE F-scores also show the same pattern for the ICSI data. However, the differences are less clear for the AMI data where the range of scores is much more compressed. In order to understand the relationship between the generated summaries and differences in standard measures, we look at the distribution of extracted content over meeting as well as summary redundancy. We find that summaries based on dialogue act level prosody better reflect the amount of human annotated summary content in meeting segments, while summaries derived from prosodically augmented lexical features exhibit less redundancy.

Index Terms: meeting summarization, prosody, dialogue.

1. Introduction

Automatic analysis of meetings has become an increasingly important task as more and more meetings are held online, recorded and archived. For example, automatic meeting summarization is a potentially useful tool for browsing and analysing dialogues for post-meeting tasks such as decision audits [1]. Methods for this have generally focused on lexical content. However, information in spoken dialogue can be characterized many other ways. For example, several studies have shown that automatic extractive summarization is possible using only prosodic features [2, 3, 4, 5]. Performance on this task is usually evaluated based on dialogue act or n-gram matching with gold standards (e.g. ROUGE [6]). Beyond this, however, we would like to understand how different aspects of prosody relate to what goes into meeting summaries. To do this, we need an understanding of where prosodic features can be incorporated into models and how this affects the generated summaries.

Current approaches generally use prosodic features calculated over dialogue acts (DAs) and the contribution of prosody is usually treated as independent to that of lexical content. Corpus analyses have found that extracted dialogue acts are generally produced with overall 'bigger' prosody, e.g. higher mean and maximum pitch and energy [3]. However, using DA level prosodic features washes out the use of prosody in marking out specific words as being important from an information structure point of view [7, 8]. It is possible that word level prosodic measurements could be used as lexical features in a similar way to term-frequency measures, such as tf.idf, which are commonly used to measure the importance of specific words relative to overall meeting content [9]. In fact, combining word prosody with tf.idf scores has been shown to help keyword extraction from voicemail messages [10], punctuation annotation [11] and topic tracking [12] in broadcast news. As such, we would like to know how these two types of information interact and how to combine them to improve summarization and dialogue understanding in general.

In this paper, we investigate whether augmenting lexical features with prosodic information improves extractive summarization performance in meetings. Our hypothesis is that integrating prosodic information at the word level will improve extractive summarization performance over plain lexical features like tf.idf. At the first stage, we combine lexical and prosodic features using an MLP to predict whether an isolated word belongs to an Extracted Dialogue Act (EDA). The probabilities generated, our *augmented lexical features*, are fed into the higher level DA extraction task. We compare the performance of our augmented features with DA level combinations of term-frequency and prosodic features, evaluating the resulting summarizers with commonly used retrieval measures, Area Under Reciever Operating Characteristic (AUROC) and ROUGE [6]. While feature analysis in summarization generally focuses on improvements of precision/recall based scores, these measures don't tell us much about how summarizers based on different feature sets vary in what they selected as a region of interest. To start to tease some of these issues out we also look at the distribution of extracted DAs in the meeting timeline and also levels of redundancy of the different summaries.

2. Experimental Setup

2.1. Data

The experiments described in the following were carried out on the ICSI [13] and AMI [14] meeting corpora. The ICSI corpus contains recordings of 75 naturally occuring meetings drawn from 8 different ongoing research groups (3-9 speaker per meeting). We use the scenario data from the AMI meetings corpus (140 meetings). Each of these meetings involved 4 speakers who worked on designing a remote control. Each group par-

This work was supported by the European Union under the FP7 project inEvent (grant agreement 287872).

ticipated in a series of 4 meetings focusing on different stages of the design process. In the following experiments, we used standard AMI development and test sets (20 meetings, 5 groups each). For the ICSI corpus, we use the same test set as [3] and a randomly selected development set (6 meetings each).

Human extractive summaries are available for all of the ICSI meetings and for 137 of the AMI meetings. There were 6 annotators involved in creating summaries, with 2 contributing to both corpora. Annotations were based on manually segmented dialogue acts and time aligned transcriptions. Annotators selected dialogue acts with the explicit goal of helping an external stakeholder (e.g. department head) understand what happened in the meeting. There was no upper limit on how many dialogue acts annotators could select for the extractive summary, although a rough guideline of 10% was given. Where possible extracted DAs (EDAs) were then linked to statements in human authored abstractive summaries that they supported. As in [3], we focus on detecting only linked EDAs as they are more likely to be actually important for understanding what happened in the meeting.

2.2. Features

2.2.1. Prosodic Features

F0 and intensity data was extracted using Praat at 10ms intervals. For F0, parameter settings were automatically determined using the method described in [15]. To reduce pitch tracking errors we calculate these parameters over spurts: segments separated by at least 500ms silence based on word alignments [16]. Missing F0 values were obtained through linear interpolation after octave jump removal. The F0 values were speaker normalized into semitones relative to speaker mean F0 value (Hz) for that conversation. Intensity measurements were normalized by subtracting the speaker mean for the conversation. We calculate prosody statistics — mean, standard deviation, maximum and minimum over F0 and intensity — over words and DAs. Utterances often exhibit natural downdrift which means prosodic gestures that are later in an utterance are produced with lower F0. To correct for this within spurts, we first subtract the predicted values from linear regression from observed values before calculating aggregate features when the slope of the spurt is negative. We include the slope as a feature at the DA level along with the other features (DA-pros).

2.2.2. Term-Frequency Based Lexical Features

For each of the words in meeting, we calculate tf.idf and su.idf as described in [9] with standard stopwords set to zero. To represent a range of speech genres, the inverse document frequency component of these measures was calculated over the combined AMI, ICSI and TDT-2 corpora. The su.idf metric takes speaker term frequency into account and was shown to have good performance in [9]. We also calculate two Pointwise Mutual Information (PMI) features which separately measure the association between words and EDA/non-EDA status in information theoretic terms [17]. PMI values are calculated over the training set only and words that do not appear in the training set are set to zero. We are primarily interested in the relationship between prosody and tf.idf, as this metric is widely used across spoken language processing tasks other than summarization. It is also applicable in single speaker situations (unlike su.idf) and does not require labelled data to estimate (unlike PMI). All values are derived after applying the Porter stemmer. For DA level prediction, we sum individual term-frequency values over words in a given DA (e.g. DA-tf.idf). This approach has been shown to be useful in a number of studies [9, 18]. We refer to the model including all three summed features as DA-tsp.

Feature	ICSI	AMI
tf.idf	0.509	0.558
su.idf	0.513	0.517
pmi	0.567	0.669
tf.idf.pros	0.532	0.601
su.idf.pros	0.546	0.599
pmi.pros	0.562	**0.673**
pros	0.532	0.602
tsp	**0.569**	0.668
tsp.pros	0.563	0.672

Table 1: *Development set AUROC for word level EDA detection.*

2.2.3. Augmented Lexical Features

Our strategy for learning how to combine prosodic and term-frequency features at the word level was to cast the problem as classifying whether a word is in an EDA or not, using an MLP to learn feature combination weights. MLPs with one hidden layer were implemented using the theano toolkit [19] with the number of hidden states tuned on the development set. We use the probabilities determined by the top logistic regression layer as our augmented lexical features. We combined each of the three term-frequency features with all word level prosodic features ({tf.idf, su.idf, pmi}.pros). We also looked at a prosody only model (pros), a model including all of the term-frequency features (word-tsp), and finally including all features (word-tsp.pros). We compare AUROC for augmented features with logistic regression classifiers using the single term-frequency features (Section 3.1).

2.3. EDA Detection and Evaluation

We use multilevel logistic regression to examine the efficacy of different feature sets for detecting EDAs. To account for differences between annotators and meeting types, as well as the unbalanced nature of the data, we include indicators for these at the group level. We model different annotators and meeting types as being drawn from normal distributions [20]. The meeting types were indicators for the 4 remote control design stages (AMI), and the 8 research groups (ICSI). For evaluation, we present AUROC and ROUGE-1 F-scores (i.e. unigram matching) calculated with DUC standard parameters [21] and a 15% word compression rate. Previous work has argued ROUGE-1 to be the most relevant version of ROUGE for meeting summarization [22]. We use gold standard annotations for calculating AUROC. Additional annotations were used to calculate ROUGE scores. The number of annotators for test set meetings was between 3-5 for the ICSI corpus and 2-3 for the AMI corpus.

3. Results

3.1. Word Level Prediction

Table 1 shows AUROC results for detecting words in summary linked EDAs on the development sets. The results indicate that augmenting tf.idf and su.idf with prosodic features improves classification performance, with a greater improvement seen for the AMI data. In fact, the prosody only based classifier performs better than both tf.idf and su.idf for both corpora. How-

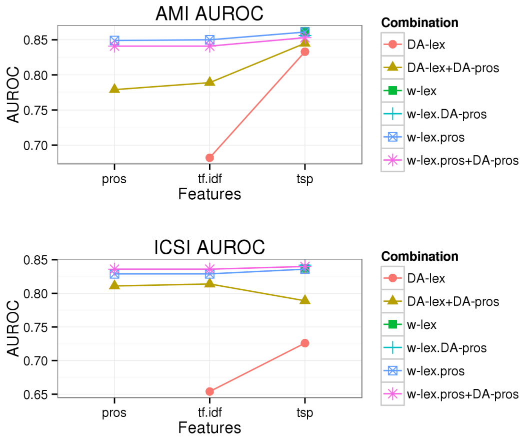

Figure 1: *AUROC for AMI and ICSI Test sets*

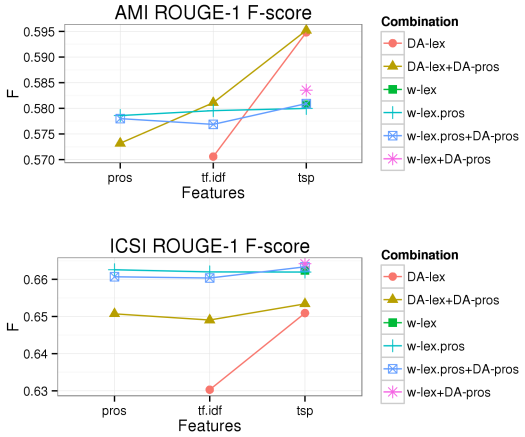

Figure 2: *ROUGE-1 for AMI and ICSI Test sets.*

ever, adding prosodic information does not appear to improve on PMI performance. Incorporating different types of lexical information (tsp) appears to help, though at this stage the results seem bounded by the performance of the domain specific PMI features. In the following, we examine the performance of these features on DA level classification.

3.2. DA Level Prediction

Figure 1 shows the AUROC results when combining term frequency features at the DA level (DA-lex), the word level (w-lex), and using our prosodically augmented lexical features (w-lex.pros).[1] We also look at the effect of adding DA level prosodic features (DA-pros). The results show that our augmented lexical features (prosody or term-frequency based) outperform DA level combinations of the same features in AUROC (w-lex.* vs DA-lex.*). The addition of DA prosody improves on bare lexical features. However, adding DA prosody to the augmented lexical features only helps for the ICSI data.

Even though DA prosody does not appear to add as much as word prosody, it's worth noting that DA prosody models are much better than reported in [3] (ICSI:0.728 vs 0.811, AMI:0.73 vs 0.799). Excluding the new group level indicators and slope features only reduces AMI AUROC to 0.778, so the majority of the gain seems likely due to improved feature extraction. In fact, contra to previous analyses, with improved F0 analysis, we find that EDAs actually have significantly lower mean pitch than non-EDAs on average, but keep expanded pitch range. We also tried including prosodic delta features, however they did not produce much of a change in performance (ICSI: 0.812, AMI:0.780, ± 4 DAs)

ICSI AUROC for augmented word-tf.idf.pros and word-tsp perform better than the best performing (full) feature set in [3] (0.829, 0.837 vs 0.818). Similarly, the word-tsp combination performs better than the best previously reported for AMI (0.861 vs 0.855). The improvement is greater for the ICSI data where term-weight features were previously reported as less effective. The augmented features perform better than DA duration (ICSI:0.813, AMI: 0.82) and length in words (ICSI: 0.831, AMI: 0.845), however the differences are quite small. Note, although PMI was a dominant predictor for the word level classification task, DA-PMI performs worse than DA-tsp (ICSI: 0.699 vs 0.726, AMI:0.798 vs 0.833), which in turn performs worse than word-tsp. In fact, incorporating prosody into word-tsp basically gives the same performance (AMI, ICSI: w-tsp=(0.862, 0.838, w-tsp.pros=0.861, 0.837). So, including extra term frequency features is more useful at the word level, however prosodic features can bridge the gap in their absence.

In summing over lexical features, we assume they weight words as being more noteworthy for summarization purposes. So, we can view our augmented lexical features improving the weighting over bare tf.idf or PMI for DA aggregation. Our augmented features provide more information than the DA length in words (a uniform weighting). While summing over such term weights has generally proven dominant in utterance level retrieval statistics like AUROC [23, 3], they have been reported as less effective when looking at n-gram based ROUGE [4, 22]. Figure 2 shows ROUGE-1 results for the different features sets. The ROUGE-1 scores mirror AUROC results for the ICSI data, with scores for the augmented lexical features significantly higher than DA-tf.idf. However, the AMI data is less clear. Although DA-tf.idf still has the lowest score, the best performance for this data set comes from DA-tsp (0.595) though the range of differences are all within bootstrap confidence intervals. Note that ROUGE scores are several points lower than for ICSI although AMI AUROC is similar or higher. Given these differences in rankings, we would like to know how these differences manifest in actual generated summaries. The next sections look at some other ways to measure summary differences.

3.3. Redundancy

While ROUGE scores give us an indication of how much overlap there is with a human summary, it doesn't tell us much about redundancy, although this often appears as a constraint in unsupervised approaches [24, 22]. We measure summary redundancy by holding out each DA and measuring its cosine distance to the rest of the summary and sum the distances [24]. Figure 3 shows a similar pattern to what we saw for AUROC measures. Summaries based on augmented lexical features were all significantly less redundant than those based on bare lexical features with and without DA level prosody (Wilcoxon $p < 0.01$, Holm corrected). Note, ROUGE-1 *recall* scores were the same or better for the augmented lexical features than DA-level prosodic features. So, while the feature based classification method doesn't explicitly take redundancy into account, longer dialogue acts tend to contain more independent information and potentially provide better lexical coverage.

[1] For *pros*, w-lex.pros includes just the prosodic features over words.

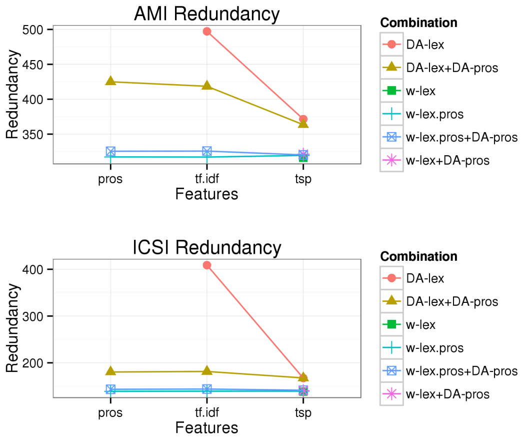

Figure 3: *Summed redundancy.*

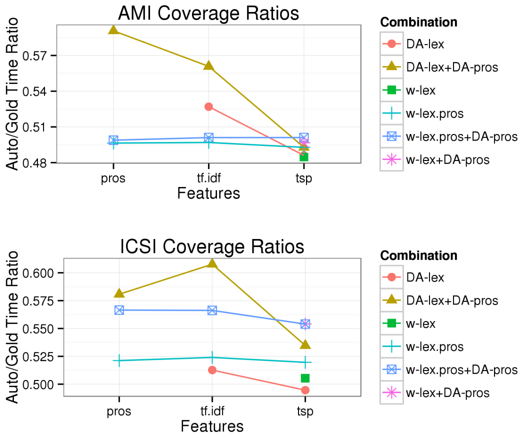

Figure 4: *Ratio of EDA time to gold standard EDA time in meeting quarters.*

3.4. Distribution of EDAs

Beyond DA content, we are interested in the temporal location of noteworthy parts of a dialogue for understanding meeting timelines. To see if the generated summaries reflect this we look at the proportion of summed EDA time from the summaries to that of the gold standard calculated over meeting quarters. Figure 4 shows that the highest average proportion arises from DA level prosody models. Although DA level prosody seems to lead to a higher proportion for both test sets, the differences are not significant (Wilcoxon signed-rank test). Nevertheless, examining the types of DAs selected by prosody models could still be helpful for improving summarization. Even though roughly the same number of words is selected in summaries for each meeting, the DA prosody model generally selects shorter DAs on average (6.12s versus 9.84s for word-tsp). So, by picking up more short utterances, prosody based classifiers may be good indicators of regions of interest, even if selection of EDAs is not as precise.

4. Discussion

AUROC ranks models in a consistent way over both data sets in our experiments, with our prosodically augmented lexical features performing better than bare lexical features with or without DA level prosody. While this provides a better weighting when summing over words in a DA, it also biases selection towards long DAs. This strategy misses out on potentially interesting types of summary information. 50% of ICSI EDAs have less than 14 words, while the mean selected by the word-tsp model is 32. DA level prosodic features are less tied to utterance length and DAs selected using this feature set were indeed shorter on average. However, adding DA prosody on top of augmented features did not really weaken the length bias. In this vein, [4] found that prosodic models gained higher weights in decision level combination with non-prosodic models in their best models. It would be beneficial to investigate in more detail how more structured higher level combination of DA level features effects retrieval performance.

Of course, we cannot expect retrieval measures to tell us everything we want to know about summaries. While ROUGE style metrics present a way of measuring whether summary relevant concepts are covered, they don't differentiate lexical content from different parts of the meeting. In general, understanding what prosodic models select beyond ROUGE style content matching is important for improving meeting summarization beyond isolated DA extraction. Although our DA prosody models did not have as high AUROC or ROUGE-1 F-scores as the text based features, they may be good indicators of regions of interest for browsing purposes. Identifying such regions rather than DAs may also help incorporate temporal aspects of meetings into automatic abstractive summaries. Predicting the amount of summary content could be useful for analyzing other aspects of group communication such as meeting productivity [25]. The proportional measure we presented for looking at the distribution of predicted EDAs is just a starting point and a more detailed analysis is still required. Similarly, while reducing summary redundancy is theoretically good for improving meeting coverage lexically, its utility in meeting browsing requires further user based testing.

5. Conclusion and Future Work

Our experiments indicate that augmenting lexical features such as tf.idf can improve extractive summarization and that incorporating prosodic information at the word level provides performance than using DA level features in AUROC terms. We found that summaries derived from prosodically augmented lexical features exhibit less redundancy. While DA prosody generally performed worse in retrieval terms, it may provide important information for temporally locating larger regions of interest. Although the objective measures discussed above present different perspectives on meeting summaries, understanding what they mean requires more extrinsic evaluation.

Our next steps will embed summaries into meeting browsing tasks in order to shed light on whether notions like redundancy or the temporal distribution EDAs affect user efficiency and satisfaction. We are also working on inclusion of features related to visual emphasis. Given differences between ICSI and AMI ROUGE-1 results, the effect of different meeting structures between corpora also requires further investigation. Similarly, since AUROC and ROUGE-1 rankings only matched for the ICSI data, so it would be interesting to see how applying compression techniques as in [26] change ROUGE scores. Another approach is to incorporate prosodic features in unsupervised summarization methods that more closely fit ROUGE's objectives [22]. Similarly, it may be useful to look at keyword identification as an objective for the generation of augmented lexical features, an approach initially explored in [10].

6. References

[1] G. Murray, T. Kleinbauer, P. Poller, T. Becker, S. Renals, and J. Kilgour, "Extrinsic summarization evaluation: A Decision Audit Task," *ACM Transactions on Speech and Language Processing*, vol. 6, no. 2, pp. 1–29, Oct. 2009.

[2] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Interspeech 2005*, 2005.

[3] G. Murray, "Using Speech-Specific Characteristics for Automatic Speech Summarization," Ph.D. dissertation, University of Edinburgh, 2008.

[4] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, Dec. 2009, pp. 387–391.

[5] S. Jauhar, Y. Chen, and F. Metze, "Prosody-Based Unsupervised Speech Summarization with Two-Layer Mutually Reinforced Random Walk," in *IJCNLP 2013*, 2013.

[6] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of Text Summarization Branches Out*, no. 1, 2004.

[7] R. Silipo and F. Crestani, "Prosodic stress and topic detection in spoken sentences," in *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*. IEEE Comput. Soc, 2000, pp. 243–252.

[8] S. Calhoun, "The theme/rheme distinction: Accent type or relative prominence?" *Journal of Phonetics*, vol. 40, no. 2, pp. 329–349, 2012.

[9] G. Murray and S. Renals, "Term-weighting for summarization of multi-party spoken dialogues," in *Machine Learning for Multimodal Interaction IV*, vol. 4892, 2007.

[10] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, vol. 2, no. 1, pp. 1–24, 2005.

[11] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

[12] C. Guinaudeau and J. Hirschberg, "Accounting for prosodic information to improve ASR-based topic tracking for TV Broadcast News," in *Interspeech 2011*, 2011, pp. 1401–1404.

[13] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 1, 2003, pp. I–364.

[14] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[15] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction." *Journal of the Acoustical Society of America*, vol. 128, no. 4, p. 2291, 2010.

[16] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 2003, pp. 34–36.

[17] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proceedings of EMNLP'06*, no. July, 2006, pp. 364–372.

[18] S. Xie, "Automatic extractive summarization on meeting corpus," Ph.D. dissertation, University of Texas at Dallas, 2010.

[19] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010, oral Presentation.

[20] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge, 2007.

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

[22] K. Riedhammer, B. Favre, and D. Hakkani-Tür, "Long story short Global unsupervised models for keyphrase based meeting summarization," *Speech Communication*, vol. 52, no. 10, pp. 801–815, Oct. 2010.

[23] G. Penn and X. Zhu, "A Critical Reassessment of Evaluation Baselines for Speech Summarization." in *ACL 2008*, no. June, 2008, pp. 470–478.

[24] K. Zechner, "Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, Dec. 2002.

[25] G. Murray, "Learning How Productive and Unproductive Meetings Differ," in *Canadian AI 2014*, Montreal, Canada, 2014.

[26] F. Liu and Y. Liu, "Towards Abstractive Speech Summarization: Exploring Unsupervised and Supervised Approaches for Spoken Utterance Compression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1469–1480, Jul. 2013.