AUDIO-BASED UNSUPERVISED SEGMENTATION OF MULTIPARTY DIALOGUE

Pei-Yun Hsueh

School of Informatics
University of Edinburgh
2 Buccleuh Place, Edinburgh EH8 9WL

ABSTRACT

In this paper, we explore a novel way to leverage audio information for unsupervised segmentation of multiparty dialogue. Our system which segments directly on patterns derived from audio sources is evaluated with previous work that segments on lexical patterns found in transcripts. We examine the effectiveness of both systems on recovering a two-layer structure of meeting dialogue. We demonstrate that the audio-based system performs significantly better than the word-based system on this task. In particular, it effectively recover segments of off-topic discussion. Results are encouraging as the audio information used in the system can be obtained in near real time and with absence of manual and ASR transcripts. It is particularly desirable when a system has to be operated online, or in unfamiliar domains and languages.

Index Terms— meetings, clustering methods, acoustic signal processing

1. INTRODUCTION

This paper addresses the challenge of segmenting meeting recordings directly from the inputs of its audio source. In particular, we focus on approaches that can be used to segment a meeting when still in progress, since we expect this to be important to the development of downstream online applications that require immediate content-based access. In fact, many automatic segmentation systems have been developed to structure meeting recordings into a number of coherent segments [1, 2, 3, 4, 5]. Typically, the task is decomposed into a series of binary decisions, each of which determines whether an utterance end contains a segment boundary or not. The dominant approach is to train a classifier with rich features that are obtained from both word transcripts and audio inputs. Although this approach has achieved success, it has some shortcomings. For one, training a well-performing discriminative model requires plentiful labelled data; yet, it is uncertain whether the trained model can be applied to segment meetings in a domain different from the labelled data.

One solution is to apply unsupervised approaches. Many have followed TextTiling approaches, first put forth in [6], to find optimal segmentation by locating lexical changes over meeting speech [1]. These works in unsupervised segmentation commonly assume the availability of manual transcripts or automatic speech recognition (ASR) outputs. Although word errors introduced by high-quality off-line ASR systems do not degrade segmentation performance [7, 5], we cannot assume ASR outputs of this quality to be readily available in the online scenario.

In the field of spoken language understanding, many research groups have attempted to perform segmentation without transcribing speech into word units first. Some have proposed to locate changes over acoustic units. For example, Malioutov et al. [8] use an unsupervised vocabulary acquisition technique [9] to derive sub-lexical units (i.e. those corresponding to high frequency words and phrases). So inter-utterance similarity can be used in a clustering approach, originally developed for text segmentation [10, 11]. However, it is uncertain whether the vocabulary acquisition algorithm that works in monologues (e.g., lectures) is robust to processing meeting dialogues which are recorded in a natural context. Others have proposed to locate changes in speaker activity, which are characterized by features obtained directly from audio inputs [12, 1, 5].

In this paper, we perform unsupervised segmentation over audio inputs. Our system leverages information that can be obtained from audio inputs in near real time. In Section 2, we describe how the speaker activity-enhanced phonetic representations are processed and how the changes in repetitions of phonemes and that of speaker activities are located. In Section 4, we compare our audio-based system against the system which segments meeting dialogue as text.

2. METHODOLOGY

In this work, our system find segmentation in phonetic units, which have been used as proxies of words in many spoken language understanding applications successfully. We modify LCSeg, a lexical chain-based approach proposed in [1], to segment multiparty discourse by locating dramatic changes in the phonetic units over utterances.

2.1. Phonetic Transcription

To characterize what has been transpired in a meeting, we first have to convert speech signals into a sequence of units. Previous work often do this using an ASR system. As we would like to explore the use of a more language- and speaker-independent way for such conversion, in this work we leverage a phoneme recognition model [13] that have been successfully applied to cross-language tasks, such as automatic language identification [14], and other spoken language understanding tasks, such as speech recognition and keyword spotting. The phoneme recognizer is trained on ten hours of the SpeechDat-E corpus [1], which consists of recorded spontaneous telephone conversations of 1,000 Hungarian speakers and their pronunciation lexicon [2]. Then the recognizer converts speech signals in the following three steps.

- Feature extraction: First, speech signals are divided into frames of 25 ms long with 10 ms shift. Next, for each frame the system utilizes a Mel-filter bank to obtain its short-term critical band logarithmic spectral density. Finally, temporal pattern (TRAP) feature vectors, i.e., temporal evolution of critical band spectral densities within a single critical band, are generated.

- Phoneme classification: For each critical band a neural network classifier is trained to estimate the posterior probabilities of sub-lexical classes (i.e., phonemes). Then, the outputs of these single band classifiers are merged in another neural network classifier such that a combined estimation of phoneme probabilities can be yielded.

- Representation preparation: A Viterbi decoder is used to produce phoneme strings. We then organize the sequence of phoneme strings into spurts, i.e., speaker turns with pause no longer than 0.5 seconds in-between.

2.2. Modelling Speaker Activity

Previous work has demonstrated the changes in speaker activity as indicative of multiparty discourse segment boundaries [12, 1, 5]. In this work we incorporate the following two types of speaker activity into the recognized phonetic transcripts. The first type ("SPK") includes speaker movements which are characterized by speaker noises (e.g., lip movement, cough), intermittent noises (e.g., door open, note taking), filters (e.g., 'hmm', 'ah') and pauses. The phoneme recognizer we use in this work can provide such information. The second type ("ACT") depicts how talkative each speaker is over the sequence of spurts in the phonetic transcripts. Herein speaker dominance is characterized as the number of phonemes transpired in each spurt; accordingly, we could enhance the phonetic transcription with speaker ID tags, $SPid$, each of which refers to the speaker of a recognized phoneme. Figure 1 (b) is the speaker activity-augmented version of the phoneme representation in Figure 1 (a).

(a) pau int h m o l k S spk s E m h u E k S m u: l k h E S O k S n E n spk pau int n m spk spk o m O k pau int
(b) pau int h SPb m SPb o SPb l SPb k SPb S SPb spk s SPb E SPb m SPb h SPb u SPb E SPb k SPb S SPb m SPb u: SPb l SPb k SPb h SPb E SPb S SPb O SPb k SPb S SPb n SPb E SPb n SPb spk pau int n SPb m SPb spk spk o SPb m SPb O SPb k SPb pau int

Fig. 1. Example of speaker activity-augmented phonetic representation.

3. EXPERIMENT SETUP

This paper addresses the challenge of whether we can segment a multiparty dialogue recording over its audio sources. In this paper, we perform experiments to answer the following questions: (1) Whether a lexical chain approach can be extended to find segmentation over utterances represented as phonetic strings; (2) Whether providing speaker activity information in addition to phonetic transcripts can further reduce segmentation errors; (3) Whether segmenting on these different versions of transcripts results in qualitatively different predictions.

3.1. Corpus and Annotation

In this experiment, we use a set of 48 scenario-driven meeting recordings from the AMI Meeting corpus. These recordings come with manual annotations of hierarchical structure and segment descriptions of these meeting dialogues. We follow previous work to flatten the hierarchical annotations into a two-layer structure of ground truth. We consider all the major discussion segments as the first layer (TOP) and aggregate all the segments in the annotation as the second layer (ALL). The functional segments (FUNC), which serve the purpose of smoothing the procession of a discussion rather than that of contributing to the discussion, are also labelled[3]. On average, each meeting is divided into 14 segments at the second layer (ALL), with around 8 segments at the first layer (TOP); in this two-layer structure, functional segments (FUNC) account for around 42% of the top-level segments and 26% of all segments.

3.2. Evaluation Metrics

We evaluate the success of segmentation systems using three different metrics: overall segmentation error rate (in Pk and

[1] Eastern European Speech Databases for Creation of Voice Driven Teleservices. $http://www.fee.vutbr.cz/SPEECHDAT-E/$.

[2] We use the phonotactic model that is trained on the part of Hungarian speaker data in the corpus, because this model, as shown in [14], outperforms the phonotactic models in other languages in the language identification task.

[3] Examples of functional segments include opening, closing, chitchat, and discussion about agenda and equipment issues.

WindowDiff(WD)), time-based accuracy (in precision and recall), and structural similarity between hypothesized and ground-truth segments. First, we use Pk and WD to provide an aggregated account of segmentation errors. Then, we examine which version of transcripts, among the others, yields best predictions of functional segments. We study precision, that is, the proportion of system-predicted segments which correspond correctly to at least one of the functional segments in ground truth, and recall, that is, the proportion of ground-truth functional segment boundaries which correspond to at least one of the hypothesized segments.

Finally, to understand the performance of segmentation systems in the online scenario, it is also necessary to study systems' capability on gauging the total number of segments in a target dialogue. The structural similarity score is computed as obtaining the difference between the number of system-hypothesized segments HYP_K and that of the number of reference segments in ground truth HYP_K and then dividing the difference (HYP_K-REF_K) by REF_K. The closer to zero, the more similar is the system-hypothesized segment structure to ground truth.

4. RESULTS

Table 1 demonstrates the effects of different versions of transcripts on segmentation performance. Line 1 shows the performance of the LC model, which locates changes in lexical patterns over word transcripts. Line 2-3 show the performance of the PH model, which locates changes in sublexical patterns over phonetic transcripts[4], and that of the PH+ACT model, which locates changes over speaker activity-augmented phonetic transcripts.

One important parameter to set in this unsupervised segmentation system is the number of segments. In search for segmentation systems that can work in online applications, in this experiment we perform our experiments under two conditions: in the first condition we set the number of segments as the number of reference segments (K)[5], while in the second condition we use a statistically determined threshold to select those most probable segment boundaries (unK)[6]. The first four columns illustrate the K condition. Results show that, when the number of segments is given, the LC model does perform better than the PH model. However, when patterns in speaker dominance (ACT) are jointly considered along with phonetic chains, the new PH+ACT model yields competitive performance to the LC model in the task of recovering top-level segments (TOP) in a dialogue structure.

The right six columns illustrate the unK condition wherein the number of reference segments is unknown. Comparing the results across the two conditions, K and unK, clearly shows a negative effect of the added structural uncertainty on the LC model, increasing the error rate[7] by 22% and 11% on recovering segments at the top level and at all levels respectively. In contrast, the added uncertainty does not significantly affect the performance of the PH model. For the task of recovering the top-level segments, the PH model outperforms the LC model by 10%; Adding the model of speaker dominance (PH+ACT) further reduces the error rate by 14%.

As functional segments covers nearly half of the top-level segments (see Section 3.1), we expect the accuracy of predicting functional segments to be important to the success of the models for top-level segmentation. Therefore, we perform subsequent experiments to examine the effects of speaker activity-based information on the accuracy of functional segment predictions. Line 1-3 in Table 2 show the results of operating the system on lexical transcripts (LC), phonetic transcripts (PH), and speaker activity-enhanced phonetic transcripts (PH+ACT). Line 4-5 show the results of locating changes in speaker movements and in speaker dominance respectively. Line 6 shows the result of locating changes in both of these two types of speaker activity information. Results suggest that, when the number of segments is given, all the systems that locate changes in speaker dominance patterns (i.e. ACT, PH+ACT, SPK+ACT) yield better precision and recall than LC. In the more realistic condition wherein the number of segments is unknown, these systems still yield higher precision than LC, with the expense of recall.

The columns of SSim in Table 1 and Table 2 demonstrates the level of structural similarity between the prediction of each of these systems that operate on different versions of transcripts and the ground truth. The close-to-zero figures of the predictions among ACT-related models (such as PH+ACT, ACT, and SPK+ACT) indicate that these systems are better at predicting off-topic functional segments (FUNC).

5. CONCLUSION

Many lexical and non-lexical patterns can be used to recover discourse structure in meeting recordings. Previous work in unsupervised segmentation uses only the lexical patterns obtained on word transcripts. In this work, we explored a novel way to capture lexical patterns, that is, to convert the audio inputs into a sequence of phonetic strings and to derive sub-lexical patterns therein. In addition, we also explored two ways to model non-lexical patterns that pertain to speaker activities: speaker movement (i.e., speaker and intermittent noise, filter, pause) and speaker dominance. We have performed experiments to examine the effectiveness of these different patterns, which can be derived from the audio record-

[4]The phonetic transcripts include both phonemes and information about speaker movements.

[5]We experiment with this condition because we want to compare with many of the previous work that use this setting.

[6]Our system follows previous work to select only potential boundary sites of which the posterior probability predicted by the system are above the mean minus half the standard deviation.

[7]Since the scores of Pk and WD are both aggregated measures of segmentation error rate, we report the change in only one of them, Pk.

	K				unK					
	TOP		ALL		TOP			ALL		
Error Rate/SSim	Pk	WD	Pk	WD	Pk	WD	SSim	Pk	WD	SSim
LC	0.36	0.38	0.36	0.40	0.44	0.55	1.11	0.40	0.49	0.42
PH	0.42	0.43	0.43	0.45	0.40	0.41	0.14	0.41	0.42	-0.23
PH+ACT	0.36	0.39	0.40	0.44	0.35	0.36	-0.38	0.39	0.40	-0.58

Table 1. *Effects of operating unsupervised segmentation on speaker activity-enhanced phonetic transcripts. Pk and WD are error rates of the predictions. SSim is a measure of structural similarity of the predictions in relation to ground-truth segmentation.*

	K-TOP		K-ALL		unK		
Accuracy/SSim	Prec	Recall	Prec	Recall	Prec	Recall	SSim
LC	0.29	0.75	0.23	0.78	0.16	0.83	6.14
PH	0.27	0.65	0.21	0.70	0.28	0.69	1.91
PH+ACT	0.36	0.86	0.28	0.88	0.40	0.77	0.09
SPK	0.28	0.62	0.20	0.65	0.71	0.61	-1.00
ACT	0.38	0.84	0.25	0.84	0.43	0.77	-.0.05
SPK+ACT	0.37	0.82	0.27	0.88	0.39	0.80	0.39

Table 2. *Effects of speaker-activity models on the accuracy of functional segment prediction. Under the K-TOP and K-ALL condition, the number of manually annotated segments at the TOP and ALL level are given as a constraint for selecting top K predictions from the hypothesis, whereas the number of segments is unspecified under the unK condition.*

ings real time or at least in near real time, on the task of recovering a two-layer structure of meeting dialogues.

Experiments have shown that, when all of these phonetic and speaker activity-related patterns are considered, our audio-based system can yield results comparable to those obtained by operating the system on manual transcripts. Consider a real-life scenario wherein one has missed the first part of a meeting and do not know how many topics have been discussed, our audio-based systems can significantly outperform the word-based system.

Results are encouraging as it shows that speaker activity-augmented phonetic units can serve as proxies of words in unsupervised segmentation of meeting dialogues. Our audio-based system can segment meeting dialogues in absence of manual and high quality ASR transcripts. It is desirable to the development of segmentation components that have to be operated online, or in unfamiliar domains and languages. Also, as the automatically derived dialogue structures can make up for the lack of explicit orthographic cues (e.g., story and paragraph breaks), the audio-based system is expected to be beneficial to developing the online version of many downstream spoken language understanding applications, such as anaphora resolution, information retrieval (e.g., as inputs for the TREC Spoken Document Retrieval (SDR) task), summarization, and machine translation.

6. REFERENCES

[1] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. of ACL 2003*, 2003.

[2] M. Al-Hames, A. Dielmann, D. GaticaPerez, S. Reiter, S. Renals, and D. Zhang, "Multimodal integration for meeting group action segmentation and recognition," in *Proc. of MLMI 2005*, 2005.

[3] M. Purver, K. Krding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proceedings of COLING/ACL 2006*, 2006.

[4] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic bayesian networks," *IEEE Transactions on Multimedia*, vol. 9(1), pp. 25–36, 2007.

[5] P. Hsueh and J. D. Moore, "Combining multiple knowledge sources for dialogue segmentation in multimedia archives.," in *Proceedings of the 45th Annual Meeting of the ACL*, 2007.

[6] M. Hearst, "TextTiling: Segmenting text into multiparagraph subtopic passages," *Computational Linguistics*, vol. 25(3), pp. 527–571, 1997.

[7] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees., "The trec spoken document retrieval track: A success sotry.," in *Proc. RIAO*, 2000.

[8] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound:unsupervised topic segmentation over acoustic input," in *Proceedings of ACL 2007*, 2007.

[9] A. Park and J. R. Glass, "Unsupervised word acquisition from speech using pattern discovery. in .," in *Proceedings of ICASSP*, 2006.

[10] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proceedings of the 28th Annual Meeting of the ACL*, 2001.

[11] F.Y.Y. Choi, P. Wiemer-Hastings, and J. D. Moore, "Latent semantic analysis for text segmentation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lillian Lee and Donna Harman, Eds., 2001, pp. 109–117.

[12] S. Renals and D. Ellis, "Audio information access from meeting rooms.," in *Proc. IEEE ICASSP, volume 4*, 2003, pp. 744–747.

[13] P. Schwarz, P. Matjka, and J. ernock, "Towards lower error rates in phoneme recognition," *Lecture Notes in Computer Science*, , no. 3206, pp. 465–472, 2004.

[14] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. Eurospeech2005*, 2005.