Gestural Cues for Sentence Segmentation

Jacob Eisenstein and Randall Davis

ABSTRACT

In human-human dialogues, face-to-face meetings are often preferred over phone conversations. One explanation is that non-verbal modalities such as gesture provide additional information, making communication more efficient and accurate. If so, computer processing of natural language could improve by attending to non-verbal modalities as well. We consider the problem of sentence segmentation, using hand-annotated gesture features to improve recognition. We find that gesture features correlate well with sentence boundaries, but that these features improve the overall performance of a language-only system only marginally. This finding is in line with previous research on this topic. We provide a regression analysis, revealing that for sentence boundary detection, the gestural features are largely redundant with the language model and pause features. This suggests that gestural features can still be useful when speech recognition is inaccurate.

Funding

This research is supported by the Microsoft iCampus project and the sponsors of MIT Project Oxygen.

1 INTRODUCTION

Gesture is frequently used with speech during face-to-face human communication [2]. Face-to-face meetings are typically preferred, and this may be due in part to additional information presented via gestural cues. If so, it is natural to ask whether machine understanding of human speech could also benefit by attending to gestural cues.

This paper explores the role of gesture in locating sentence unit (SU) boundaries. Prosody has already been shown to be a useful feature, improving performance beyond that achieved by language modeling alone [9, 14]. We describe a system that classifies sentence boundaries based on gestural features, and then combines the gestural and language models in a single sentence segmenter. We find that gestural cues do correlate well with sentence segmentation, but that the information provided by gesture is also highly correlated with features from the speech modality. A regression analysis reveals that the gesture *residual* – the unique information carried in the gesture track – is a weak and noisy signal. This explains why we and others [1] have not seen large performance improvements from gestural cues, and suggests that such cues might be more helpful in the context of recognized – rather than transcribed – speech.

We begin by describing some related work in the field, then describe our approach, including our corpus of multimodal data and the specific gesture features that we chose. We describe the details of our implementation, then move on to the results. The ensuing discussion section explores the reasons for the relatively poor performance of the gesture feature. Finally we conclude by proposing future work and summarizing our findings.

2 RELATED WORK

A series of papers have used small corpora to show relationships between linguistic phenomena and gesture (e.g., [5, 13]), although they do not include automatic systems for recognizing the linguistic phenomena based on gestural cues. We know of no implemented system for any natural language problem where gestural features are shown to produce a statistically significant improvement over a state-of-the-art non-gesture system. Nonetheless, the prior research does imply that such improvement is possible, and suggests which gestural features should be used.

Quek et al. describe the relationship between discourse structure and gestural cues in [13]. They find that "catchments" (defined as the repetition of two or more gesture features) are often indicators of topic shifts in the discourse structure. Esposito, McCullough, and Quek describe the relationship between filled (e.g., "ummm") and unfilled (silent) speech pauses in [5], work that could eventually improve speech recognition by using gestural cues to help detect filled pauses in speech.

Chen et al. [1] recently described an application of gesture features to sentence segmentation. This system was based on a gestural corpus of three videos, that were specifically chosen because they had a relatively low "hold" rate – in other words, the speakers gestured frequently. Using a language model trained on the CTS dataset from the Switchboard corpus [7], they trained a hidden-event language model, then tried to show improvement using prosodic [14] and gestural features. A human-corrected

computer vision system was used to obtain the gesture features semi-automatically, and a decision tree model trained on these features performs betters than chance. Chen et al. show a small improvement when adding the gesture cues to prosodic and verbal models, but this improvement was not statistically significant.

We extend these results, using a larger corpus and a wider range of gesture modeling and model combination techniques. More importantly, we use hand-annotated gestures features, following the widely-used taxonomy defined by McNeill [12]. While using hand-annotation is disadvantageous from the perspective of building an end-to-end system, it gives us access to a much wider feature set than is presently possible with automatic transcription. Even with these extensions, we still find that gesture affords only a relatively modest improvement over language and prosody. We present a regression analysis that explains this phenomenon and begins to quantify the limits of the benefits offered by the gesture modality.

3 CORPORA

This section describes two corpora – a multimodal corpus gathered by the authors, and our application of the CTS Switchboard corpus used only for training the language model. All testing used the multimodal corpus.

3.1 Multimodal Corpus

The multimodal corpus includes nine speakers – four women and five men, between the ages of 22 and 28. Eight of the participants were right-handed; eight were native English speakers. The participants ranged in age from 22 to 28. All had extensive computer experience, but none had any special experience in the task domain of explaining mechanical device behavior.

The participants were presented with three conditions, each of which involved describing the operation of a mechanical device after viewing a computer simulation. The three conditions were shown in order of increasing complexity, as measured by the number of moving parts: a latch box, a piston, and a pinball machine. In explaining the devices, the participants were allowed – but not instructed – to refer to a predrawn diagram that corresponded to the simulation. All simulations were two-dimensional, and the devices were specifically chosen to be easy to describe in two dimensions.

The explanations were videotaped and manually transcribed by the first author. Speech was transcribed with a unique timestamp for each word. Speech recognition has not been applied to this corpus, as the signal quality of the recording is too low. Gesture was transcribed using the feature set described in Section 5.

The monologues ranged from 15 to 90 seconds in length. A total of 574 gesture phrases were transcribed; as many as 58 and as few as six gesture phrases were used in a single explanation. The number of words used ranged from 25 to 270; the number of sentence units (SUs) ranged from three to 29. SU annotation was performed according to the Simple Metadata Annotation Specification [16].

An example transcript of the speech is shown below. Additional details and results from this corpus can be found in [3].

so there's a force here and it pushes down um and then as this thing comes by it goes along this little incline part here and pushes that out and that goes past it but then the spring brings it back so that this thing uh manages to latch on and the box will not open

3.2 Lexical Segmentation Corpus

The SRILM toolkit [15] contains a variety of language modelling tools, including an implementation of the lexical segmenter described in [14]. We would have liked to have trained a segmenter using our own corpus for training data, but the size of our corpus – 2698 words and 241 sentences – is too small for this purpose. The SRILM distribution includes a trigram model of sentence segmentation data, trained from the CTS Switchboard corpus, and this is used for lexical segmentation in this research. Note that there are some important differences between the Switchboard corpus and the test data drawn from the multimodal corpus. One of the most important differences is that Switchboard contains phone dialogues while the multimodal corpus contains monologues. Some of the keywords that frequently initiate sentences in the phone dialogue setting are very rare in monologue setting: for example, question words like "what" and "why." Similarly, the sentence-initializing words in the monologue setting may not serve the same role in dialogues. Consequently, the performance of the lexical segmenter was significantly worse on our corpus than documented in other studies, as described below.

4 FEATURES OF GESTURAL COMMUNICATION

This study investigates whether the presence of certain gestural features at candidate sentence boundaries is an effective supplement to a purely lexical sentence segmenter. Kendon describes a spectrum of gesturing behavior [8]. On one end are artificial and highly structured gestural languages, such as ASL. In the middle, there are artificial but culturally shared *emblems*, such as the "thumbs-up" sign. At the far end is *gesticulation*, gestures that naturally and unconsciously co-occur with speech. Gesticulation is of particular interest since it is completely natural; speakers do not need to be taught how to do it. However, gesticulation is challenging because of the potential for infinite variety in gesturing behavior across speakers.

4.1 Hierarchy of Gesture Features

Verbal communication can be described in a hierarchy extending from high-level entities that occur over relatively long periods of time (paragraphs and sentences), to intermediate-scale entities (words), on down to highly specific short-duration entities (i.e., morphemes, such as prefixes and suffixes that specify information such as verb

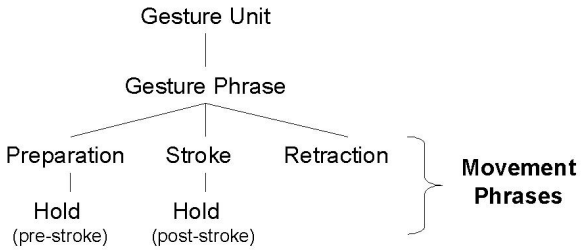

Figure 1: A hierarchy of gesture components

tense, gender, and number). Kendon describes a similar hierarchy for gestures [8], a portion of which is shown in Figure 1.

At the top level is the *gesture unit*, a period of activity demarcated by the return of the hands to a rest position. The *gesture phrase* is what is traditionally considered to be a single "gesture" – for example, pointing at something while talking about it, or tracing a path of motion. At the lowest level are *movement phrases*: morphological entities that combine to create each gesture phrase. Every gesture phrase must have a *stroke*, which is considered to be the content-carrying part of the gesture. In addition, there may also be a *prepare* phase, which initiates the gesture, and possibly a *retract* phase, bringing the hand back to rest position. A *hold* refers to a static positioning of the hand in gesture space, either before or after the stroke.

McNeill defines several types of gesture phrases [12]. In this corpus, three gesture phrase types were found to particularly relevant: deictic gestures, where the hand refers to a location in space; iconic gestures, where the form of the hand or motion depicts the action being described; and beat gestures, which convey no semantics but serve a pragmatic function, such as emphasizing certain points in the speech.

We included the pause duration feature because we are interested in the relationship between gesture and prosody. Earlier studies of prosodic cues for sentence segmentation found pause duration to be the most informative feature [14], and it is easy to implement. Our results include performance both with and without this feature, to isolate the unique contribution of gestural cues.

4.2 Validity of Annotation Features

Although this taxonomy for annotating and describing gestures is widely used both in the psycholinguistics literature and in multimodal user interfaces, there appears to be little published work validating these categories through interrater agreement. In earlier research, we found the agreement among naïve raters on gesture phrase classification was modest ($\kappa = .45$) but statistically significant ($p \ll .01$) [4].

Our experience with reliability of temporal segmentation of gestures into movement phases is similarly mixed, although it is more difficult to quantify interrater agreement on a temporal segmentation task. In the only known study of interrater agreement on movement phase segmentation, Kita reports 76% raw agreement on "gross" tem-

Name	Value	Description
Gesture Unit	Boundary, Non-boundary	A gesture unit start- or end-point occurs during the candidate sentence boundary.
Gesture Phrase	Boundary, Deictic, Iconic, Beat	Describes the gesture phrase that overlaps the candidate sentence boundary. If the value is "Boundary", then a gesture phrase begins or ends near the candidate sentence boundary.
Movement Phase	Boundary, Prepare, Stroke, Hold, Retract	Describes the movement phrase that overlaps the candidate sentence boundary. If the value is "Boundary", then a movement phrase begins or ends near the candidate sentence boundary.
Pause Duration	Real	Prosodic feature describing the duration of the candidate sentence boundary.

Table 1: Gestural and prosodic features used for sentence segmentation

poral segmentation, which falls to 72% when phase type is considered [10]. However, since it is unclear how to compute the chance agreement, these numbers cannot be converted into Kappa scores – the standard way of assessing interrater agreement on classification tasks. By examining the movement phase labels on a frame-by-frame basis, we were able to compute the probability of chance agreement. We found raw agreement of 63% ($\kappa = .45$).

5 IMPLEMENTATION

There are three components to our implementation: the language model, the gesture model, and model combination algorithm.

5.1 Language Model

As mentioned above, our language model is an identical replication of the HMM-based lexical model described in [1, 14] and implemented in the SRILM toolkit [15], using only lexical features. When pause features are included, they are handled by the same model as the gesture features. As a baseline, we also measure the performance of the language model and the pause duration without gesture features; in this case, a mixture of Gaussians is used to model pause duration.

5.2 Gesture Model

For the gesture model, we treated every space between two words as a potential sentence boundary, and built a feature vector of the gestural annotations present over this

temporal interval. The gesture features that we considered are shown in Table 1. Dividing the data into test and training sets, we applied various supervised learning techniques to learn gesture models. In earlier work on multimodal integration, decision trees were chosen for both gestures [1] and prosody [9, 14]. We experimented with decision trees, naïve Bayesian models, and AdaBoost [6]. Support vector machines and log-linear models were found to be too time-inefficient for this problem. In all cases, the Weka machine learning toolkit was used [17].

Since the number of positive instances – sentence boundaries – is greatly outnumbered by the number of negative instances, the decision tree and several other classifiers will classify all instances as negative. To address this, the training set is "balanced" so that the positive instances are weighted more heavily.

All classifiers performed significantly better when the posterior probability is compared to a learned threshold, rather than using 0.5 as the default threshold. This applies to the baseline language model classifier as well, which performed 11% better with thresholding. The ideal threshold was usually around .2, although this varied widely depending on the error metric that was to be optimized; the number cited is for the widely-used Slot Error Rate, described in the next section. Cross-validation was also employed to find a margin around potential sentence boundaries in which gesture events that occurred nearby but not exactly at the boundary would be included in the feature vector. However, the results were not very sensitive to this feature and a margin of zero was used.

5.3 Model Combination

In their research on combining prosodic features with language models, Shriberg et al. discuss three possible methods for combining models [14]. *Interpolated combination* simply adds together the output of each model, scaled by a weighting parameter that is estimated from cross-validation. *Joint-classifier combination* uses the language model posterior as a feature in a higher-level classifier that also considers the non-verbal features. *Direct modeling* integrates all features into a single HMM, where the non-verbal features are treated as emissions from the states, and the state transitions are given by the language model.

From the literature, there is no clear winner among these choices. For example, Shriberg et al. find that direct modeling outperforms interpolated combination when transcribed words are used, but that the reverse is true when the corpus consists of recognized words [14]. Kim et al. evaluate all three techniques, and find that the winner varies depending on the what is being recognized [9]. In all cases, the differences were quite small. We have implemented interpolated and joint-classifier combination.

6 EVALUATION

Evaluation was based on the Slot Error Rate (SER) metric. SER is defined by the sum of the false positives and false negatives, divided by the number of actual sentence boundaries in the document (given by the sum of true positives and false negatives).

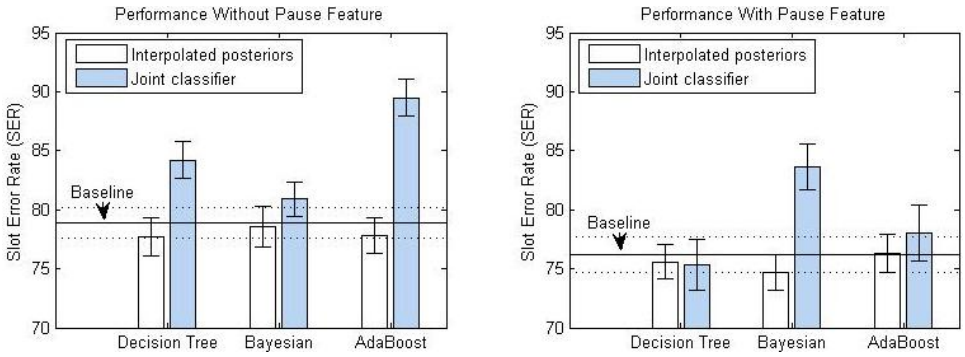

Figure 2: Performance of multimodal sentence unit boundary detection. Error bars are 95% confidence intervals.

$$SER = \frac{FP + FN}{TP + FN} \tag{1}$$

Note that this metric can exceed 1, as the number of potential false positives is bounded only by the number of words. All results are based on randomly selecting 20% of the monologues to be held out as a test set, and iterating this procedure 100 times. We hold out entire monologues rather than parts of monologues to ensure that the test and training set are truly independent. We believe this is preferable to the evaluation performed in [1], where 10-fold cross-validation is performed across only three different documents, meaning that data from at least one document is shared between the test and training sets. A still more stringent evaluation would be to require that no *speaker* is present in both the test and training sets, eliminating the possibility of speaker-specific adaptations. In future work, we will explore this possibility to determine the extent to which speaker-specific adaptations are present.

The results are shown in Figure 2. The chart on the left shows performance without the pause duration feature. The solid line across each chart is the baseline performance, without considering gesture features. The dotted lines are the 95% confidence intervals for the baseline. The gesture models improve performance in many cases, but never by a statistically significant margin.

6.1 Model Combination

The interpolation feature combination method outperforms the joint-classifier combination method almost every time. This difference is statistically significant with at least $p < .05$ $(t_{(198)} = 2.04)$ in all three cases when the pause feature is absent; with the pause feature present, the difference is again significant for the Bayesian classifier at $p < 1 * 10^{-10} (t_{(198)} = 7.08)$. The difference is not significant for AdaBoost with the pause feature, and the joint-classifier technique actually performs slightly better than the interpolated classifier technique with decision trees when pause features are present.

6.2 Gesture Model

The picture revealed by our data is somewhat less clear on the choice of gesture models. Without the pause feature, the Bayesian model performs a little worse than the other two classifiers; with the gesture model it outperforms them. None of these differences were statistically significant at even $p < .1$. The fact that there was no clear winner even after averaging over 100 trials suggests that the choice of gesture model will not significantly impact performance, and that the choice of classifier should be made on other criteria. The Bayesian classifier ran slightly faster than the decision tree and a good deal faster than AdaBoost, but the decision tree can claim an advantage in human readability.

6.3 Features

In the absence of a gesture model, the pause feature significantly improved performance over lexical features alone ($p < .01, t_{(198)} = 2.68$). The pause feature also significantly improved performance over the LM+Gesture model combination with the Bayesian classifier ($p < .001, t_{(198)} = 3.59$), and trended towards improved performance with the decision tree ($p = .058, t_{(198)} = 1.91$) and AdaBoost ($p = .18, t_{(198)} = 1.34$).

While gesture features show a trend towards improved performance, the gains were not statistically significant. Without pause features, the decision tree was the best gesture model classifier, but the 1.2% difference in SER was not statistically significant ($p = .245, t_{(198)} = 1.17$). With pause features, the Bayesian gesture model offered a raw difference of 1.5% SER, compared to a combination of the language model and a Gaussian model of the pause duration, but this was not statistically significant ($p = .167, t_{(198)} = 1.39$). These results mirror the findings in [1], where the gesture model afforded only a small improvement, within the margin of error.

7 DISCUSSION

The error rates reported here are somewhat higher than those given in [1]. In both studies, the Switchboard CTS corpus is used to train the language model, but where Chen et al. use dialogues in their test set, we have monologues. As discussed above, there are important differences between these types of speech, and it is not surprising that using incompatible test and training sets diminishes performance of the language model.

Gesture patterns also differ significantly between monologues and dialogues [3]. In a monologue, there is a greater prevalence of semantic gestures such as iconics and deictics; discourse-moderating beat gestures that control turn-taking and provide backchannel feedback are less common. It is possible that these gestures are more relevant to sentence segmentation, reducing the performance of the gesture model on monologues.

Table 2 shows a Bayesian gesture model that was learned from a subset of our corpus. Without the help of a language model, this model classifies sentence boundaries

Feature	Value	$p(.\|S_{boundary})$	$p(.\|\neg S_{boundary})$
Gesture Unit Boundary	TRUE	.039	.0069
Gesture Phrase	BOUNDARY	.21	.088
	DEICTIC	.30	.43
	ICONIC	.46	.46
	BEAT	.031	.028
Movement Phase	BOUNDARY	.27	.13
	PREPARE	.066	.057
	STROKE	.31	.46
	HOLD	.26	.30
	RETRACT	.090	.052

Table 2: Bayesian model of gesture data

correctly with an F-measure of .28. Due to a very large number of insertion errors, the SER of this model alone is 1.81, but its performance is still significantly better than chance. Note the differences in the conditional probabilities of the gesture features based on the sentence boundary status. Several of the gesture features appear to correlate well with the sentence segmentation information.

If the gestural features are contributing unique information, then the fault probably lies with our model combination techniques, and we should look there to improve performance. However, there is also the possibility that the gestural cues are better correlated with the LM posteriors than with the sentence boundaries themselves. This view has a basis in the psycholinguistics literature – Krauss argues that the gesture modality is not an independent channel for the expression of semantics, but rather, is derivative of the speech channel and cannot provide any new information about the underlying semantics [11]. If this is the case, there is no model combination technique or gesture model classifier that could substantially improve performance.

7.1 Regression Analysis

To assess the level of interdependency between the gesture, pause, and language models, we performed a multivariate linear regression. The results of this analysis are shown in Table 3. The first column of this table shows the linear correlation between each model and the true sentence boundaries. While each correlation is statistically significant, the lexical features are the most informative; a linear transformation of these features can be used to explain $.42^2 = 18\%$ of the variance of the sentence boundaries.

We then performed a multivariate regression to extract the residual of each model with respect to the other two models. The residual is the variance in each model that is not explained by a linear combination of the other two models. This captures the ability of each model to make a *unique* contribution to the sentence segmentation. The second column shows the correlation between the residual of each model and the true boundaries. Note that not only is the lexical model most informative, but it carries the highest proportion of unique information. 74% of the information carried by this model remains in the residual – as opposed to 34% for the pause data and only 9% for the ges-

Feature	r_{model}	$r_{residual}$	$r^2_{residual}/r^2_{model}$
Lexical	.42	.36	.74
Pause	.29	.16	.34
Gesture	.17	.05	.087

Table 3: Regression analysis of each feature type

ture model. The residual of the gesture model now explains only 0.25% of the variance in the sentence boundaries. This relationship is still statistically significant since the sample size is very large ($p < .03, df = 2103$), but unlikely to improve sentence segmentation in all but a very few cases. From this analysis, it is unsurprising that gestural cues did not improve sentence segmentation performance – they reinforced information conveyed by other models but contributed almost no information of their own.

8 CONCLUSIONS

The results presented in this paper show that although gesture contains information relevant to sentence segmentation, that information is largely redundant with other modalities. This phenomenon appears to be robust to the choice of gesture modeling technique, and also to the model combination algorithm.

In the presence of noisy speech or prosody data, it is still possible that gesture could improve sentence segmentation performance significantly. Another potential area of future research is whether gesture can improve the recognition of sentence-internal metadata, such as disfluencies, filled pauses, and repetitions. We have also gathered a new corpus of 90 dialogues across fifteen pairs of speakers. Using this corpus, we can investigate the differences in gesture and speech between monologues and dialogues. Finally, a study to establish whether gestural cues affect the sentence segmentation decisions of human raters would help us to better set our expectations for the role of gesture in automatic sentence segmentation.

8.1 Acknowledgements

For helpful insights and conversations about this work, we thank Aaron Adler, Christine Alvarado, Regina Barzilay, Sonya Cates, Lei Chen, Lisa Guttentag, Tracy Hammond, Mary Harper, Michael Oltmans, and Metin Sezgin. This research is supported by the Microsoft iCampus project and the sponsors of MIT Project Oxygen.

References

[1] L. Chen, Y. Liu, M. P. Harper, and E. Shriberg. Multimodal model integration for sentence unit detection. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'04)*. ACM Press, 2004.

[2] N. Chovil. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194, 1992.

[3] J. Eisenstein and R. Davis. Natural gesture in descriptive monologues. In *UIST'03 Supplemental Proceedings*, pages 69–70. ACM Press, 2003.

[4] J. Eisenstein and R. Davis. Visual and linguistic information in gesture classification. In *Proceedings of International Conference on Multimodal Interfaces(ICMI'04)*. ACM Press, 2004.

[5] A. Esposito, K. E. McCullough, and F. Quek. Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE Workshop on Cues in Communication*, 2001.

[6] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning (ICML'96)*, 1996.

[7] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (Vol. 1)*, pages 517–520, 1992.

[8] A. Kendon. *Conducting Interaction*. Cambridge University Press, 1990.

[9] J. Kim, S. E. Schwarm, and M. Osterdorf. Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL'04*. ACL Press, 2004.

[10] S. Kita, I. van Gijn, and H. van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and Sign Language in Human-Computer Interaction (GW'97)*, volume 1371, pages 23–35. Spring-Verlag, 1997.

[11] R. Krauss. Why do we gesture when we speak? *Current Directions in Pschological Science*, 7(54-59), 2001.

[12] D. McNeill. *Hand and Mind*. The University of Chicago Press, 1992.

[13] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, pages 171–193, 2002.

[14] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32, 2000.

[15] A. Stolcke. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (Volume 2)*, pages 901–904, 2002.

[16] S. Strassel. Simple metadata annotation specification (version 5.0). Technical report, Linguistic Data Consortium, 2003.

[17] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.