Combining Acoustics, Content and Interaction Features to Find Hot Spots in Meetings

AUTHORS: DAVE MAKHERVAKS
WILLIAM HINTHORN
DIMITRIOS DIMITRIADIS
ANDREAS STOLCKE

What Are Hot Spots?

“ Hot spots are parts in conversations that stand out from the rest of the conversation in that:

- Participants are more involved (emotionally or ‘interactively’)
- There is a higher degree of interaction between participants who are trying to get the floor “

-- Wrede et al. [2005/ICSI Technical Report]

Why Are Hot Spots Useful?

- Improve summarization
- Support meeting analytics
- Increase human productivity

Roadmap of Presentation

- Overview of ICSI corpus
- Kornel Laskowski's paper
- Task definition
- Speech features
- Word embeddings
- OpenSmile
- Results
- Conclusions

ICSI Corpus - Overview

- 75 meetings
- 72 hours
- Average of 6 speakers / meeting
- Janin et al. [2003/ICASSP]

ICSI Corpus - Annotations

Hot spot annotations:

- 3 levels: lukewarm, warm, hot
- 3 degrees: -, 0, +
- Type: Amusement, Clarification, Disagreement, Agreement, etc.
- Labels are at the utterance level, based on linguistic segmentations

Other annotations:

- Dialog Acts
- Adjacency Pairs
- Error Codes
- Etc.

Time marks for transcribed words, with speaker labels.

- Determined by forced alignment of human transcripts on close-talking microphones

Defining a Machine Learning Task

The problem:

- Unbalanced dataset: ~1% were hot spots
- Annotations were too granular

Solution:

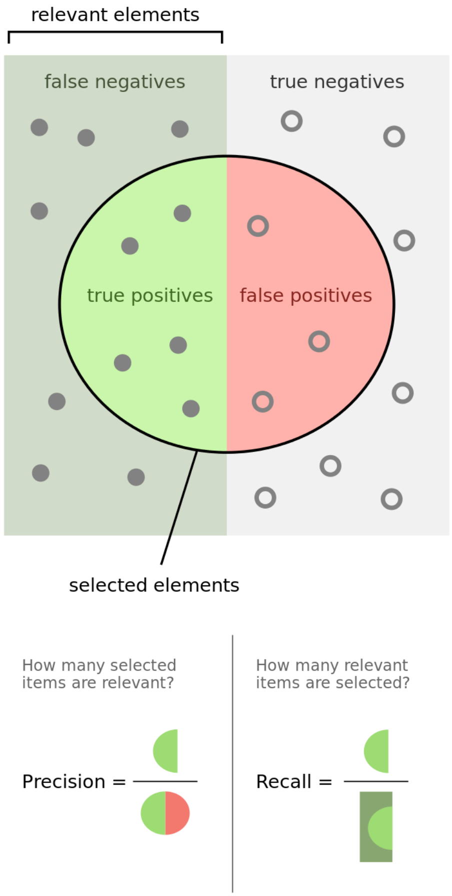
- Turn this into a binary classification problem (hot or not)
- Use uniform intervals as units

Why do we like UAR?

- UAR = unweighted average recall
- Equal to accuracy with same aggregate weight given to all classes (regardless of corpus frequency)
- Metric does not depend on class prior distribution

Metrics for Classification

- ACC – Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- UAR – Unweighted Average Recall
  - Same as ACC, but after rebalancing frequency
- Baseline for UAR: 0.5 (chance performance)



Kornel Laskowski [2008/SLT]

Key aspects:

- Detect whether a 60 second interval contains involved speech, with a 15 second shift
- Laughter is most important feature
- Only other features used: speech activity (by speaker)

)ELING VOCAL INTERACTION FOR TEXT-INDEPENDENT
)N OF INVOLVEMENT HOTSPOTS IN MULTI-PARTY MEETING:

Kornel Laskowski

: Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
ognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany

ABSTRACT

nmarization in recordings of meetings ıly on the propositional content of what objectively relevant, such content may main aim of potential system users. In- ted in information bearing on conversa- tomatic detection of one example of such hotspots defined in terms of participant l system relies exclusively on low-level yields a classification accuracy of 84%, on of error relative to a baseline which

ı processing, Meetings, Pattern classifi-

ıTRODUCTION

marization in recordings of meetings and focused largely on the propositional con- ıy. Although objectively relevant, such or even the main aim of potential system be interested in information bearing on

degree of interaction by participants who are trying to g Although it was shown in [4] that VERSION2 hots temporal extent is a function of involved utterance dur sociated with the degree of simultaneous vocalization f participants (overlap), no evidence was presented to s observed differences are discriminative.

Our objective in the current work is to present a bas detector. Using the extensive annotation of VERSION volvement (described in Section 2), but with the le temporal support of VERSION1 hotspots, we propo which classifies 60-second intervals of meetings as eith involved speech ($\mathcal{I}$) or not containing involved speec which relies only on very low-level vocal interactio might be available from a vocal activity detector. These described in Section 3, and the experiments presented and 5 demonstrate that laughter is almost solely res our reduction in error of 39.2% relative to a majority line. Section 6 compares automatic versus human and the impact of our results is briefly discussed in S summarized in Section 8.

2. DATA

By the numbers:

- 84.0% accuracy (not UAR) with laughter related features
- Laughter is a cheating feature

Train/dev/eval split – 75 total:

- 49/11/15

Revised Task Definition

Our adjustment:

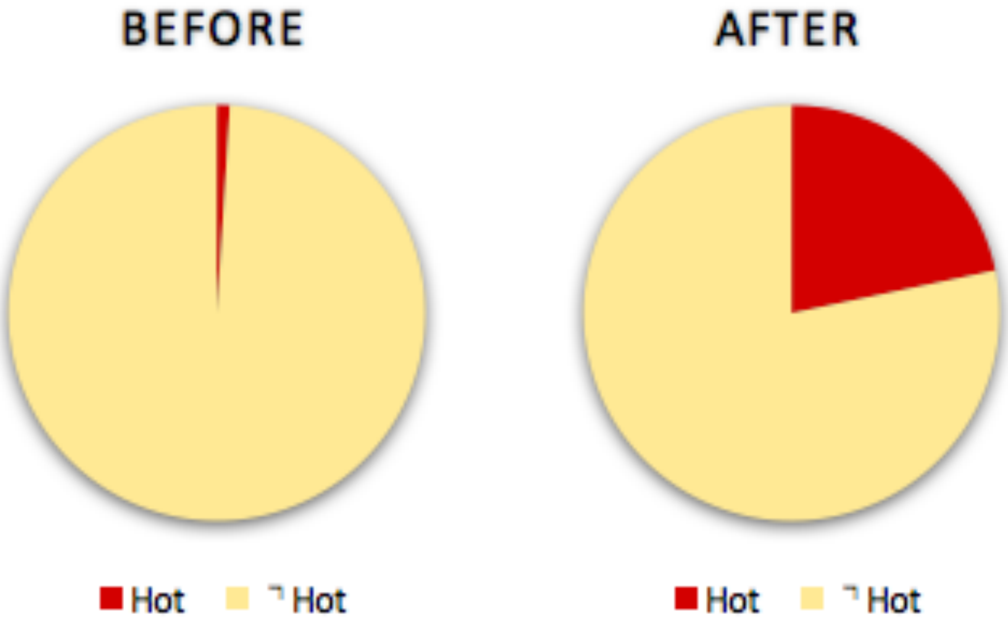
- If utterance == hot, 60 sec window = hot

Improvement:

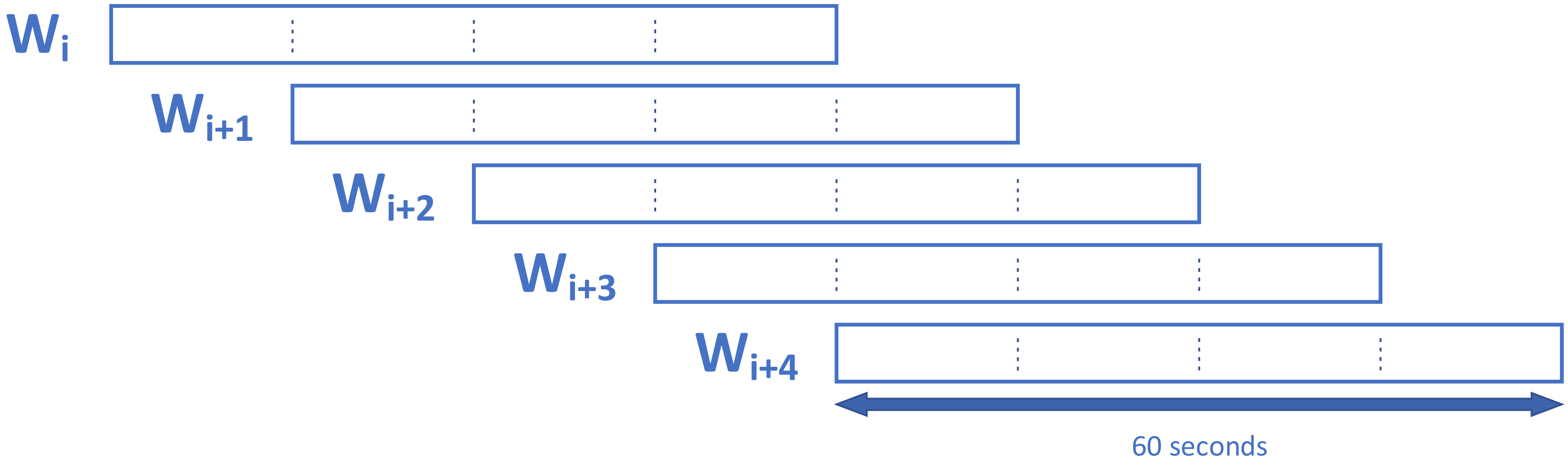
- ~22X more of minority set

Different from Laskwoski:

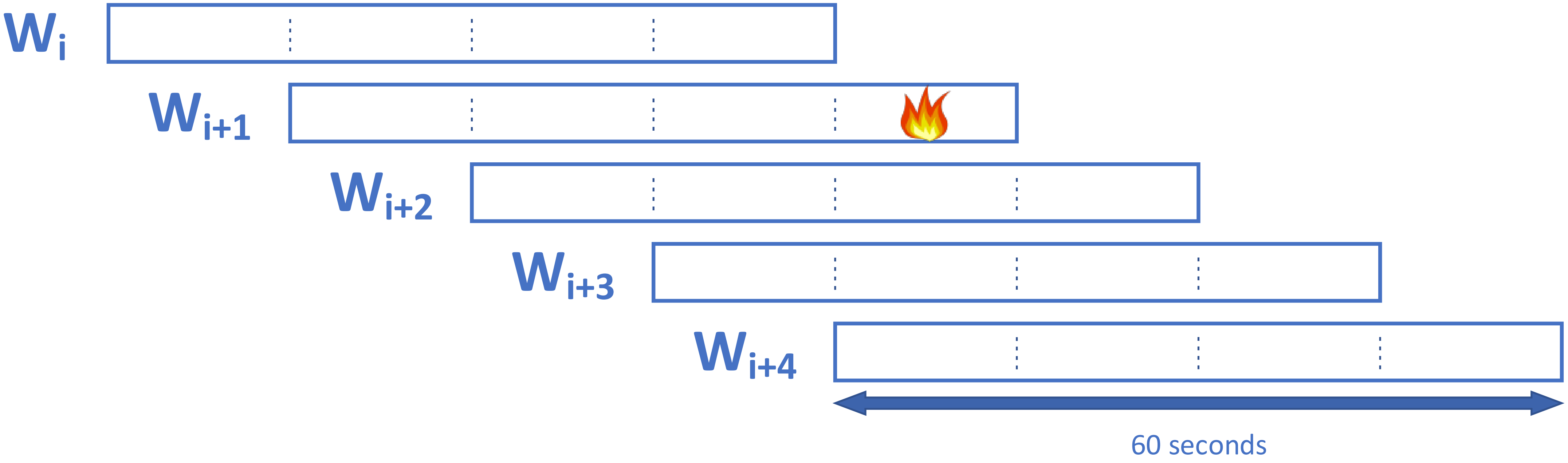
- 15649 intervals vs. 15823
- 26.6% involved vs. 21.7% involved



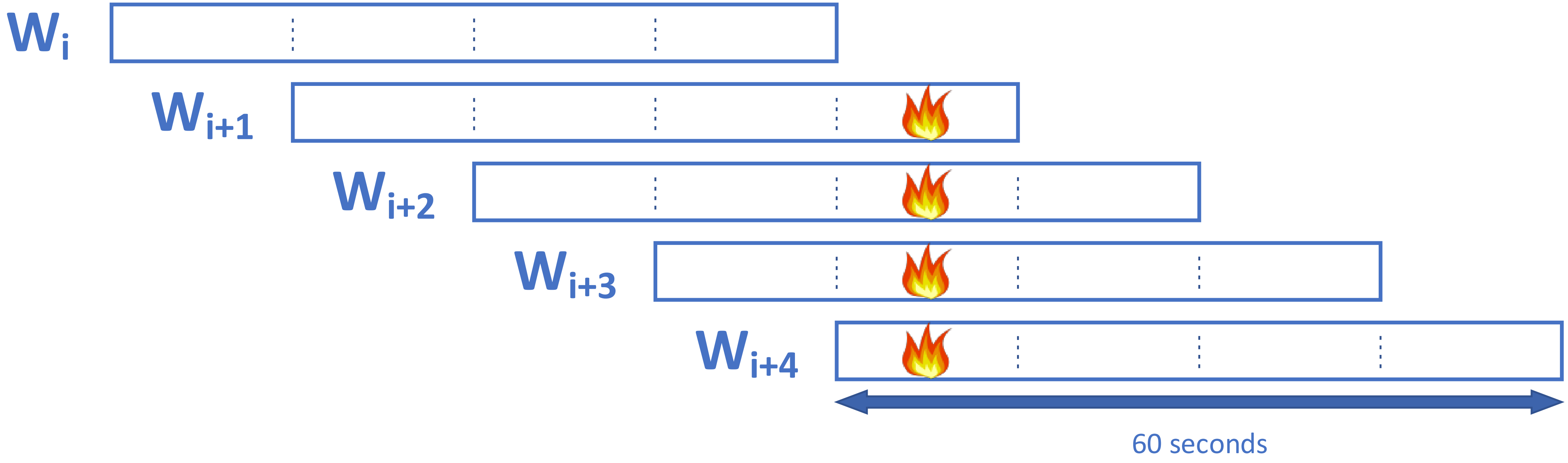
Window Visualization



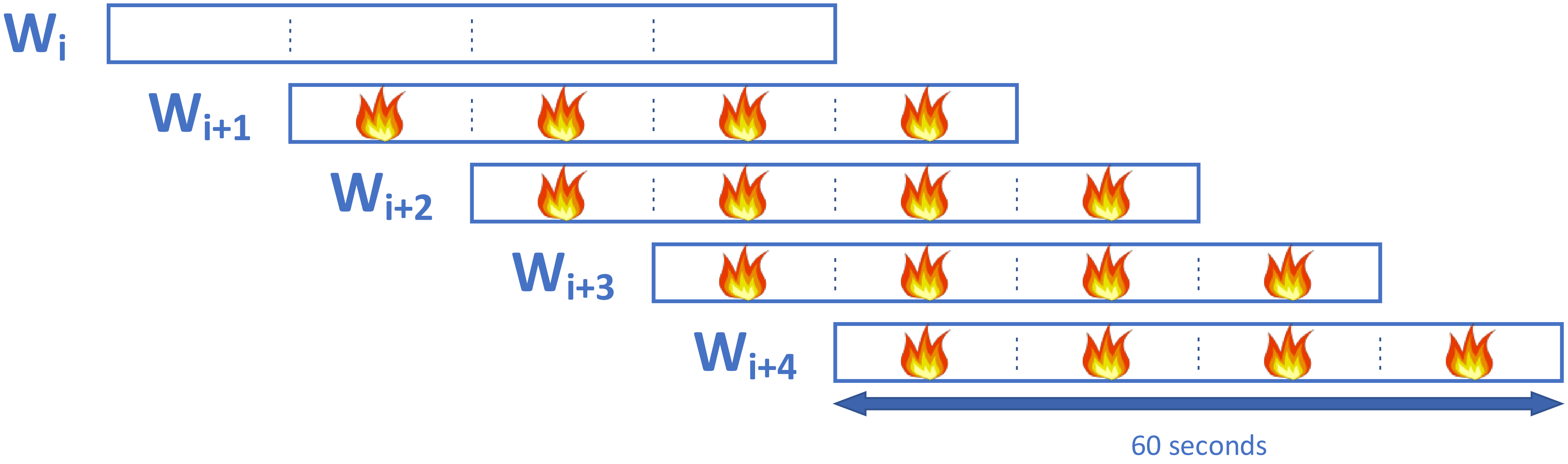
Window Visualization



Window Visualization



Window Visualization



Experiments with Speech Activity and Interaction Features

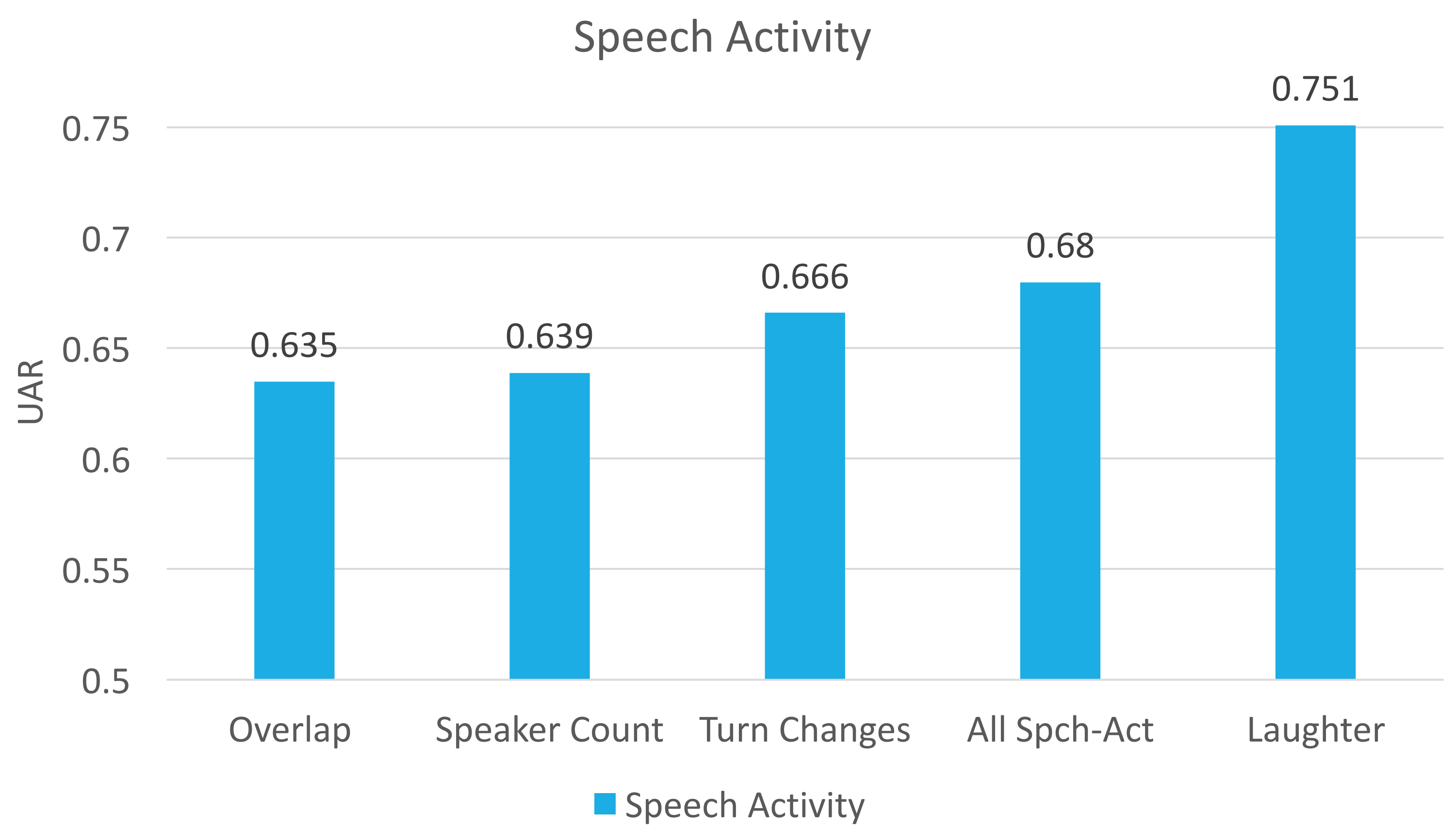
Features extracted:

- Speaker overlap percentages
- Unique speaker count
- Turn switch count

Models used:

- Logistic regression (class-balanced weight)
- Random Forests
- Multinomial Naïve Bayes (Multinomial NB)
- Linear Support Vector Machine (Linear SVM)

Speech Activity Features: Results



Laughter is great, but ...

Laughter Count

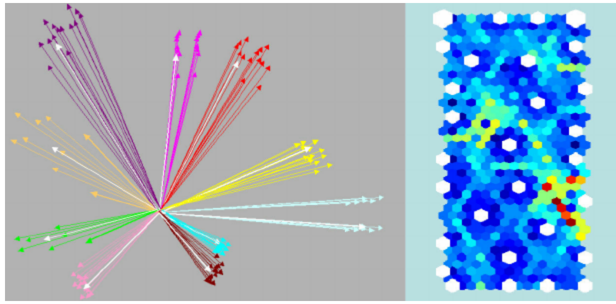
- Further research required for automatic detection
- Depends on social environment
- Network would learn to rely on laughter

Word Embeddings

Extracted from BERT model

Smaller, better representation:

- 1024 dimensions
- Proximity between embeddings = semantic & linguistic similarity



Adapted vs. unadapted

- Adapted on spoken call center corpus - used for sentiment classification
- Adapted performs slightly better than unadapted

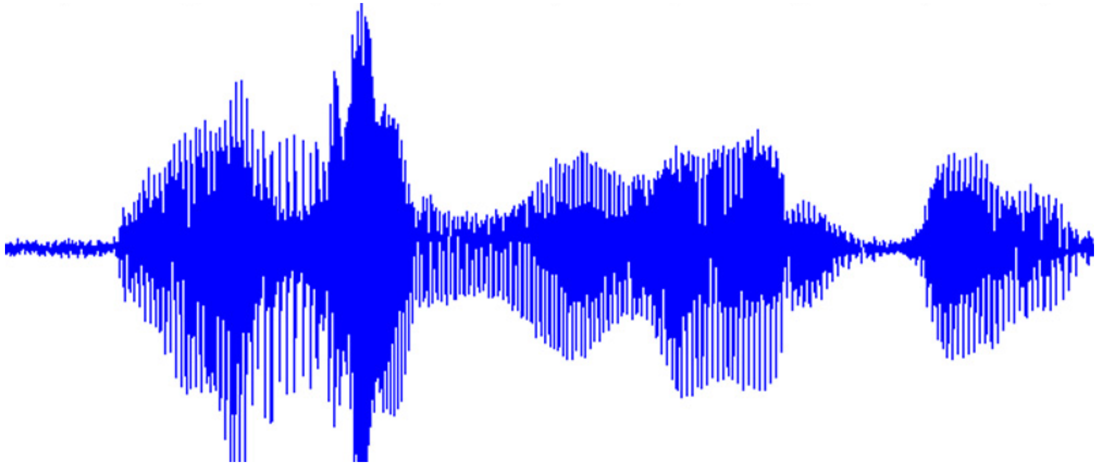
The embedding vectors are pooled over the entire window, zero-centered, and then classified

Prosodic Features

Prosody denotes the supra-segmental (above the phone level) aspects of speech that are encoded by pitch, energy, and duration

Why would they help?

- Prosody conveys emphasis, sentiment, and emotion
- Expect higher involvement to be correlated with increased sentiment, emphasis, and emotion



OpenSMILE

Standard toolkit for emotion extraction from speech

- Uses acoustic features

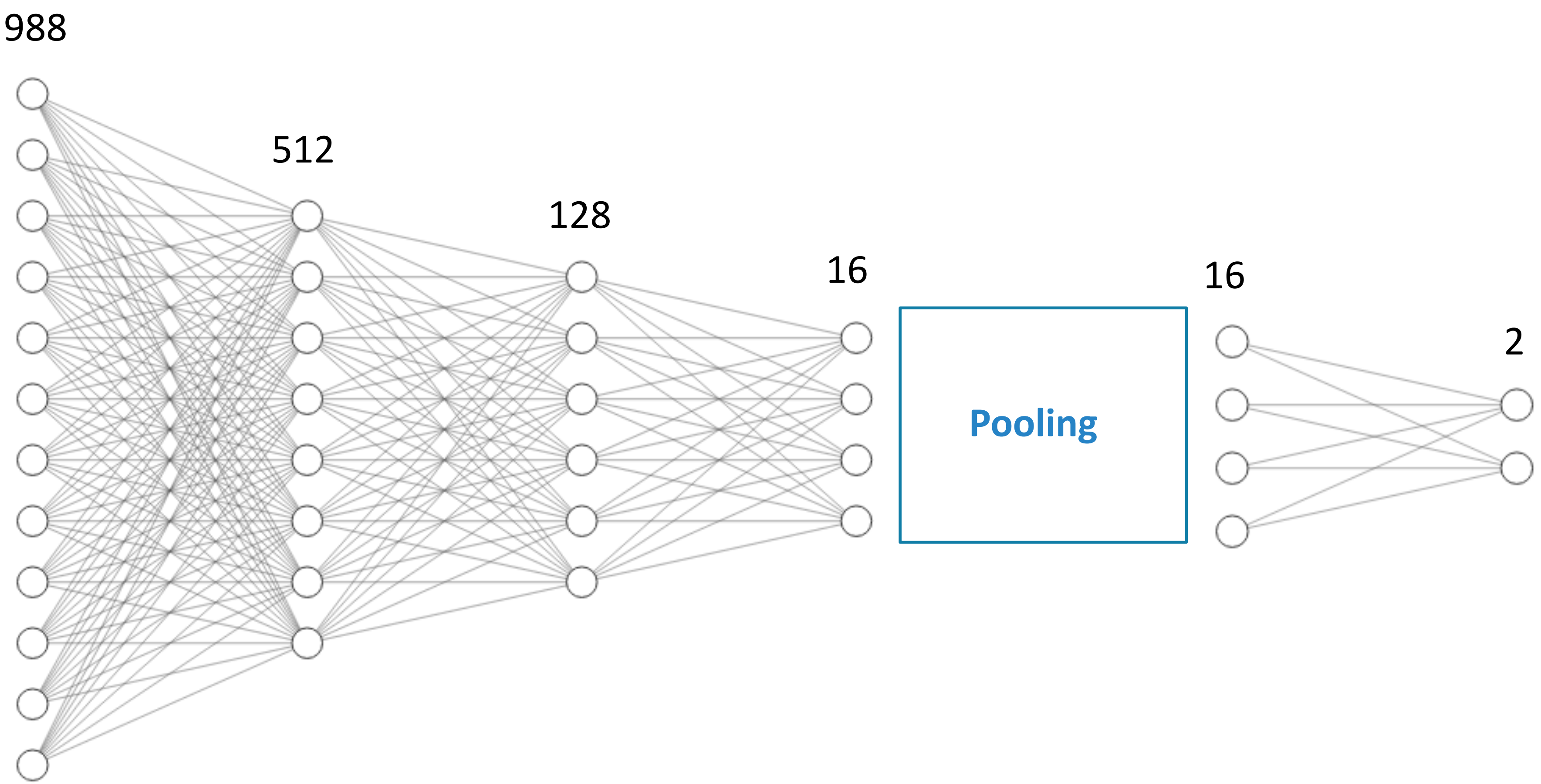
Config file used: emobase

- Helpful for emotion, sentiment detection
- 988 features

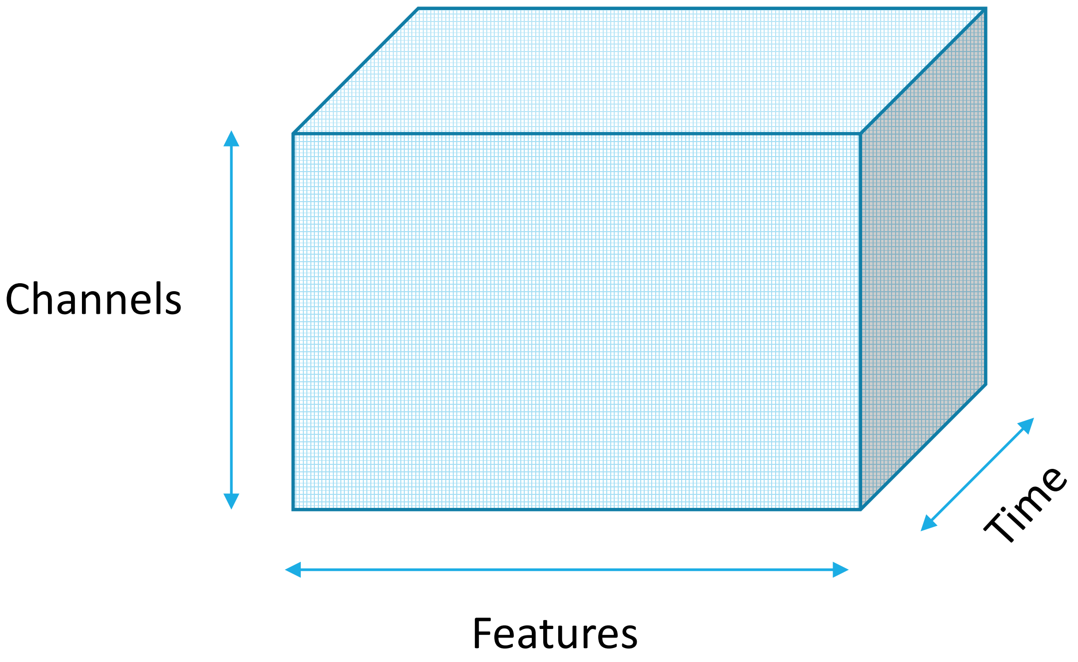
2 choices of feature extraction windows

- Entire 60 second window
- 5 second sub-windows, pooled over the 60 second window
  - OpenSmile features are designed to operate on single utterances

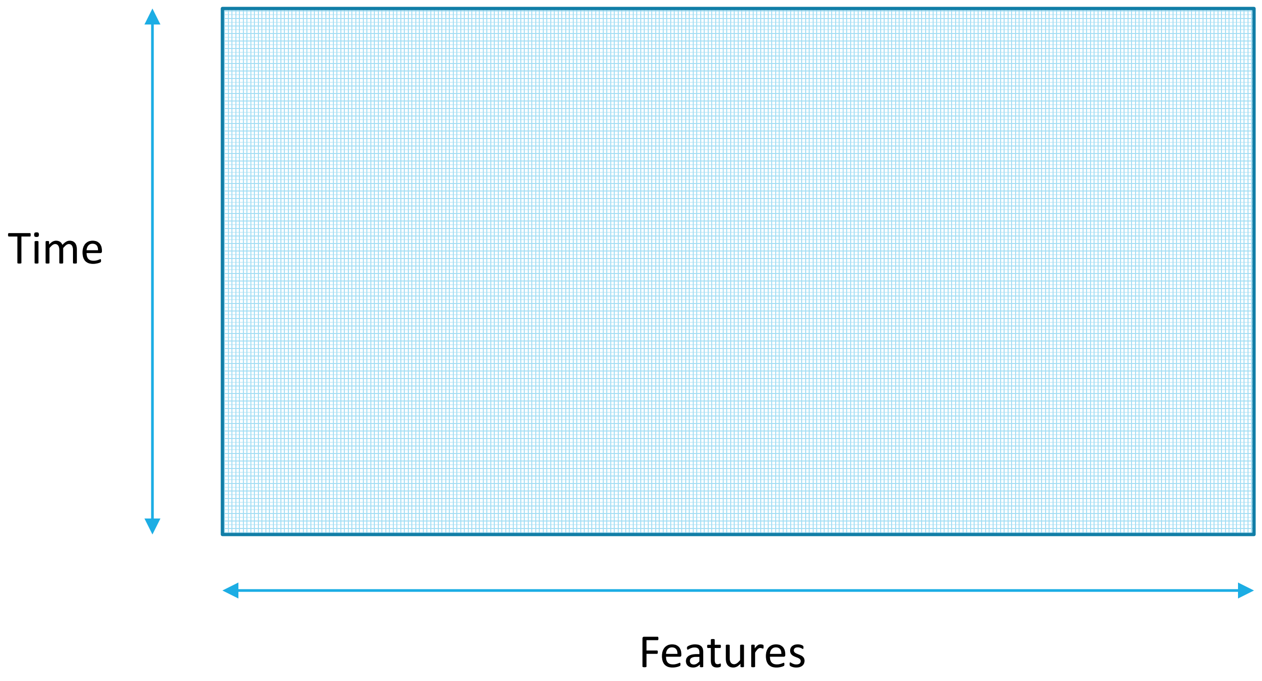
Neural Network for OpenSmile Feature Classification



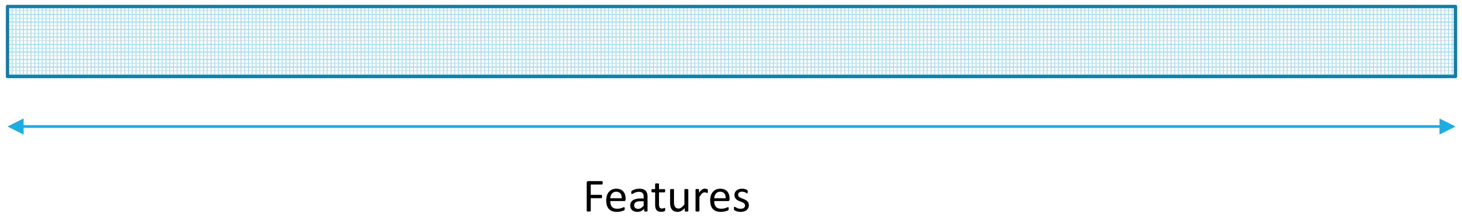
Representation of Prosodic Features



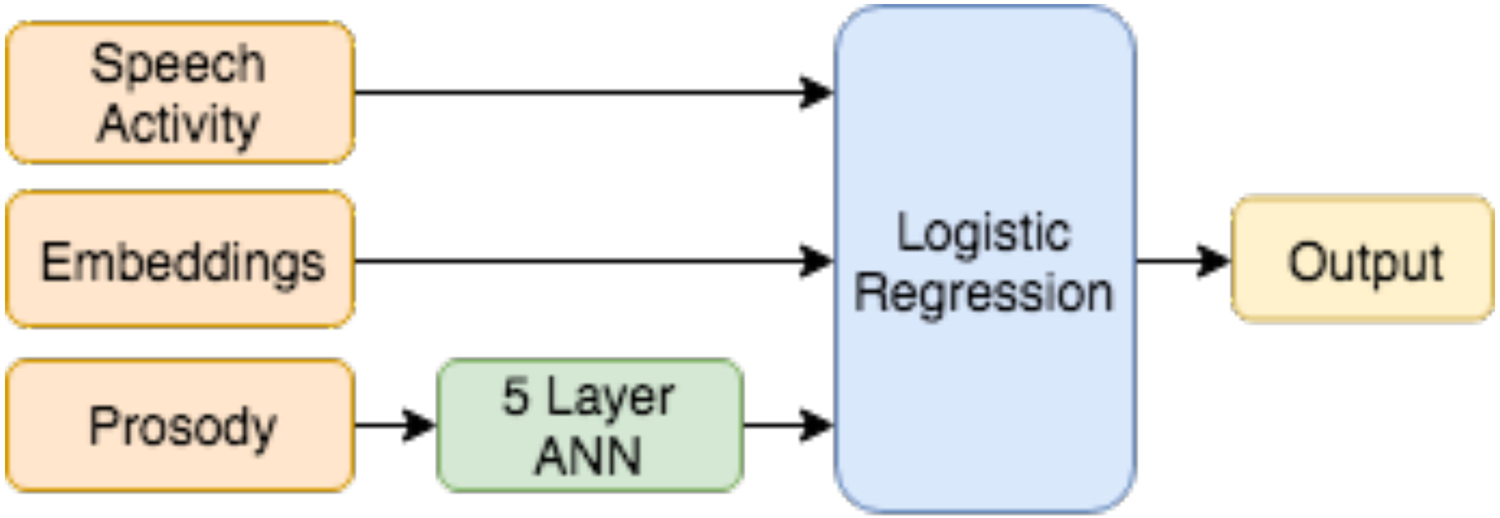
Max-pool over Channels



Mean Pool Over Time



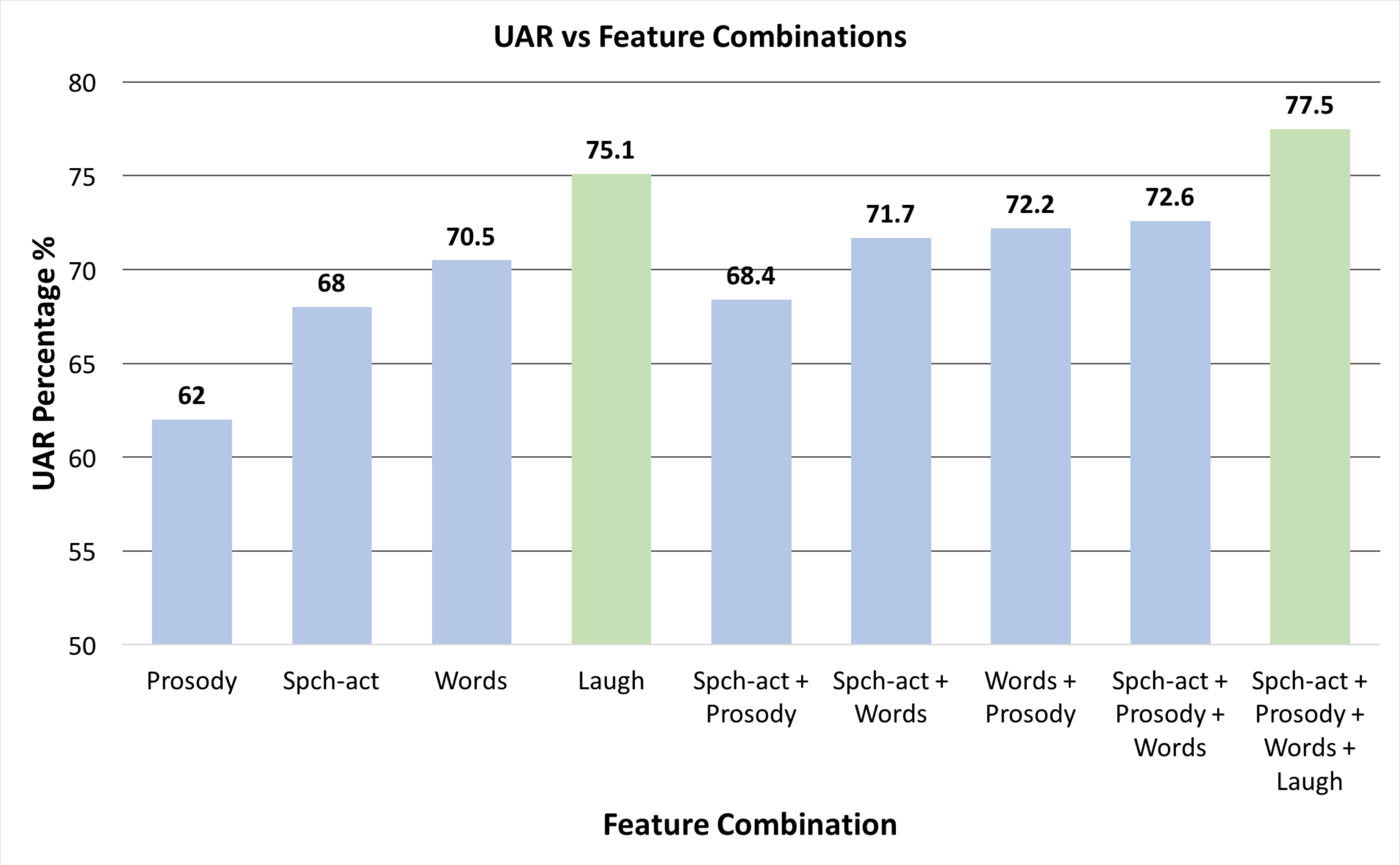
Overall Classifier Architecture



Classification Results by Feature Type

Feature Set	UAR with Features	UAR without Features
Prosody (OpenSMILE)	62.0%	71.7%
Speech activity	68.0%	72.2%
Words (BERT)	70.5%	68.4%
All	72.6%	N/A

Results with Feature Type Combinations



Hot Spot Detection: Conclusions

- A combination of word-based, prosodic, and interaction features can predict high involvement (or “hotness”) in 60-sec windows with about 73% UAR (where chance is 50%)
- Word-based features using BERT embeddings are the single most important speech-based source of information
- Prosody, while not as strong by itself, is the next most informative speech feature (in combination with words)
- Interaction features (which are based only on speech activity) are informative by themselves (as observed by Laskowski), but do not add much information once words and prosody are given
- Laughter is a very strong indicator of involvement by itself in the ICSI corpus (75% UAR), but we don’t trust that it can be extracted reliably or that it will generalize across different types of meetings.

Future Work

- Validation on other meeting corpora
- Feature extraction with automatic speech recognition
- Feature fusion by NN as opposed to Logistic Regression
- Demonstrate utility of hot spot detection in an actual meeting summarization system

Acknowledgments

- Britta Wrede
- Elizabeth Shriberg
- Kornel Laskowski