

SPEAKER CHANGE DETECTION USING A NEW WEIGHTED DISTANCE MEASURE

Soonil Kwon, Shrikanth Narayanan

Department of Electrical Engineering,
University of Southern California

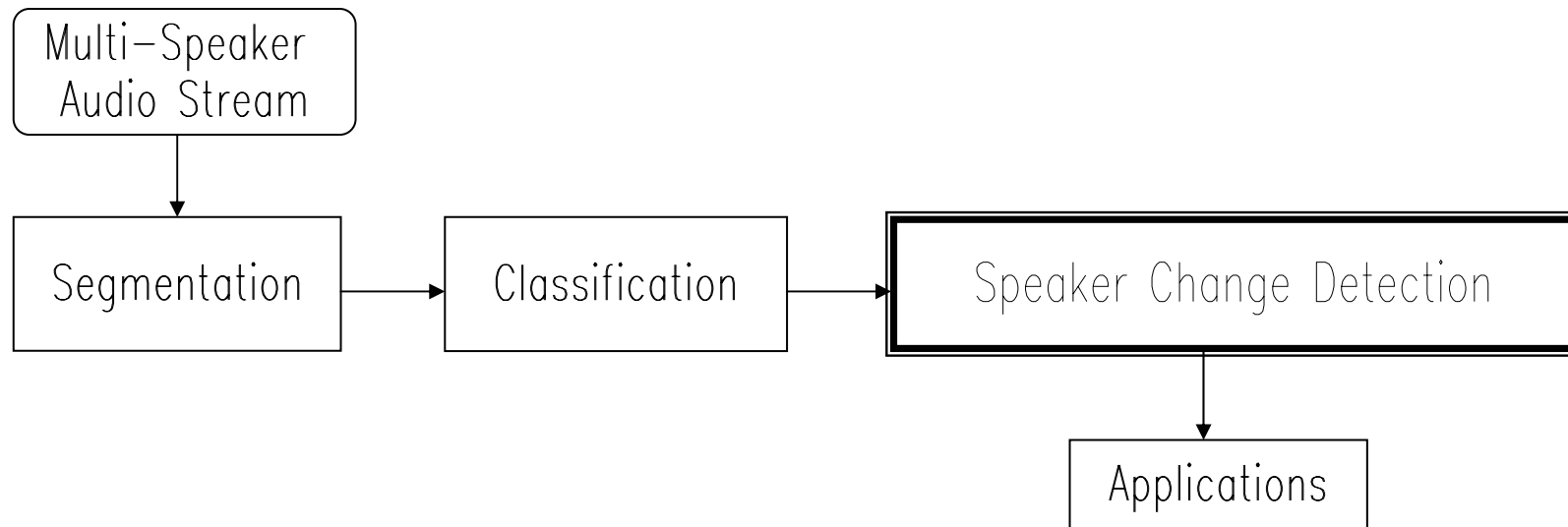


Motivation

- Model-based algorithm needs a large training data set and prior knowledge of speakers.
- Metric-based algorithm using Euclidean or Mahalanobis is simple, but not under the influence of interclass information.
- Our algorithm can be useful in speaker tracking, speaker adaptation, automatic transcription, and audio indexing system.



Speaker Change Detection





Segmentation

- To get a sequence of discrete utterances from an audio stream
- Fixed length segmentation
 - Short segment: not enough speaker information.
 - Long segment: likely to include changing points in the middle of segments.
- Variable length segmentation
 - Breathing points: speaker changes are unlikely to be between breathing points.



Classification

- Classify speech segments and non-speech segments with
 - silence ratio.
 - level of variation in the zero crossing rates.



Speaker change detection

- The identity and number of speakers are unknown.
- Detect speaker changes using ‘Weighted Squared Euclidean Distance Measure’.
- Thresholds are determined by using a small number of the data segments.



Weighted Squared Euclidean Distance

- Sample space: $X(k) = \{x_1^{(k)}, x_2^{(k)}, \dots, x_{n_k}^{(k)}\}$,
(n_k : the number of samples in class k).
- $x_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{im}^{(k)})$ is an i -th feature vector ($i = 1, 2, \dots, n_k$) where $x_{ij}^{(k)}$ is the j -th feature in the i -th sample ($j = 1, 2, \dots, m$).
- Mean vector of class k :
$$\bar{x}_j^{(k)} = (\sum_{i=1}^{n_k} x_{ij}^{(k)}) / n_k$$
- Weighted Squared Euclidean Distance

$$d_{ki} = \sum_{j=1}^m w_j^{(k)} (\bar{x}_j^{(k)} - \bar{x}_j^{(i)})^2$$



New Weights (1)

- $$\mathbf{w}'_j = \left(\frac{\mathbf{tvar}_j}{\mathbf{wvar}_j} \right) / \left(\sum_{j=1}^m \frac{\mathbf{tvar}_j}{\mathbf{wvar}_j} \right)$$

\mathbf{tvar}_j : the variance of total j -th feature vectors from two neighboring segments.

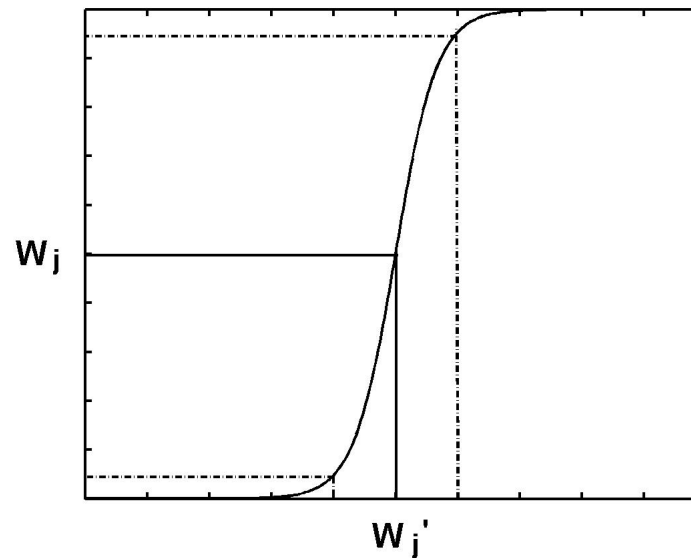
\mathbf{wvar}_j : the sum of the variance of j -th feature vectors from segment 1 and the variance of j -th feature vectors from segment 2.

New Weights (2)

- Sigmoidal Function

$$w_j = \frac{1}{1 + e^{-p(w_j' - \bar{w})}}$$

$$\text{where } \bar{w} = \frac{\sum_{j=1}^m w_j'}{m}$$



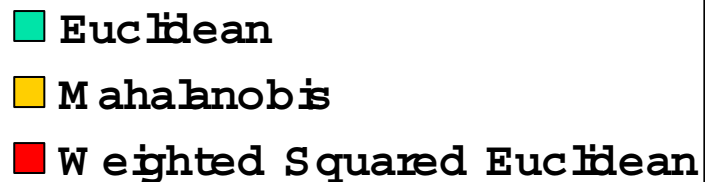
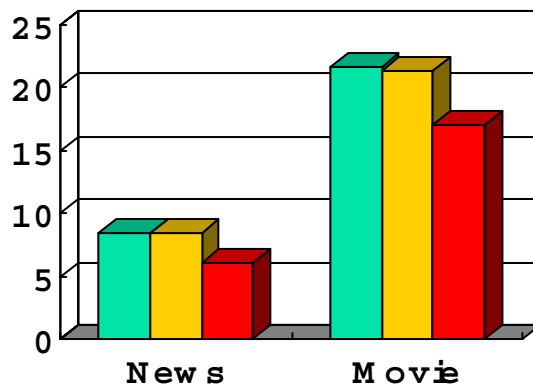


Experiments

- Sixty minutes of the broadcast news data extracted from HUB-4 Broadcast News Evaluation English Test Material(1999).
- Twenty minute movie audio data from ‘When Harry met Sally (1989)’.
- Sample: 16kHz, 16 Bit, Mono.
- Feature vector: 48 channel, 32 dimensional MFCC.
- Training data: 5 minute news, 10 minute movie audio (arbitrarily chosen).

Results

**Error rate (%) of
Speaker Change
Detection**



- Error = (false acceptance error) + (false rejection error).
- New metric provided about 37.7% improvement for broadcast news and 27.1% for movie data compared with Euclidean distance .



Discussion

- Errors of speaker change detection were due to
 1. Short speech segments.
 2. Changes of background noise within a segment.
- Errors of segmentation were due to
 1. Interruption of another speakers in the middle of speaking.
 2. Background noise in the middle of a silence region separating two different speakers.



Conclusion

- Our weighted distance measure discriminated the speakers more precisely without any knowledge of speakers and a large training data set.
- To improve the overall performance
 1. More robust segmentation algorithm.
 2. Simple, fast, but more robust detection algorithm.



Future Work

- More robust speech detection system that incorporates segmentation with classification.
- Manipulation of data that result from speaker change detection.