An Approach to Sequential Grouping in Cochannel Speech

Ke Hu and **DeLiang Wang**

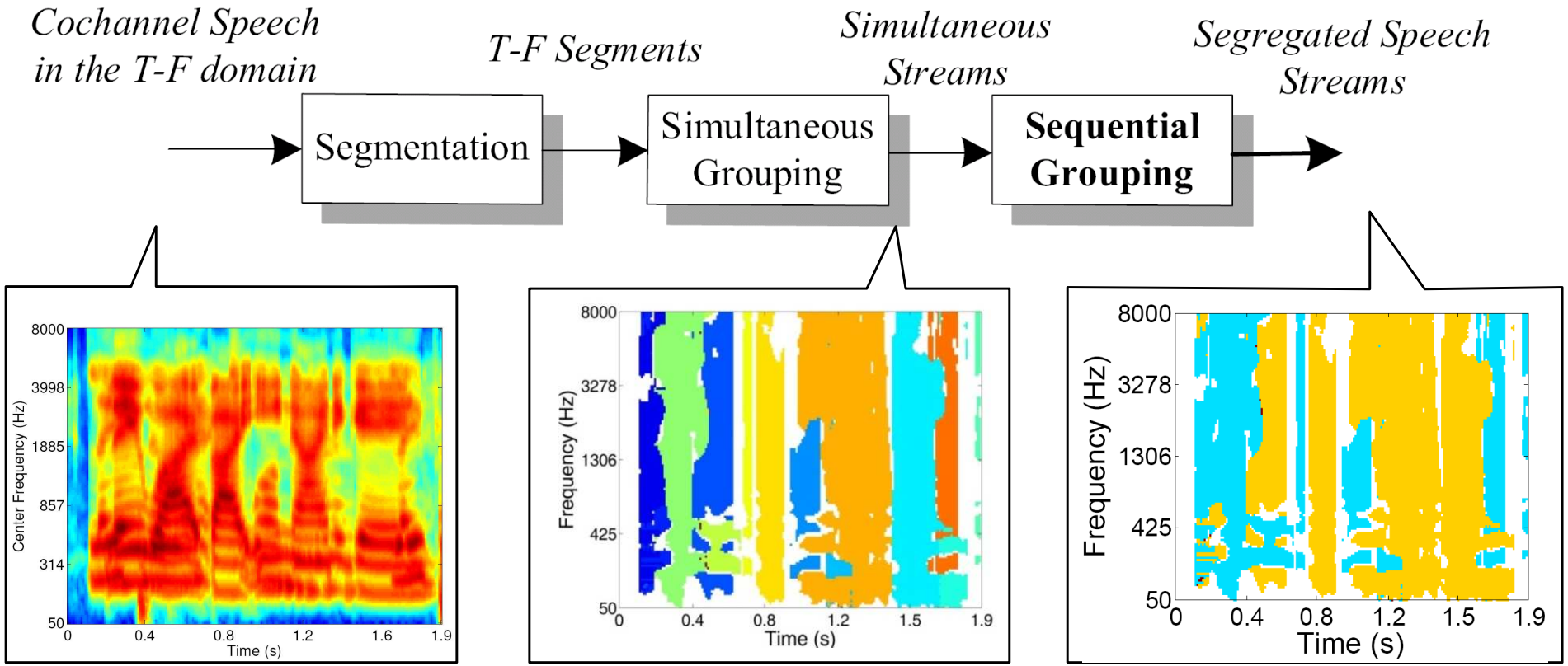
Perception & Neurodynamics Lab
The Ohio State University

Outline

- **Background**
- **Unsupervised sequential grouping**
  - Sequential grouping of voiced speech
  - Sequential grouping of unvoiced speech
- **Evaluation results**

Sequential organization problem

- **Sequential grouping in computational auditory scene analysis (CASA) aims to organize sound across time into different source groups**

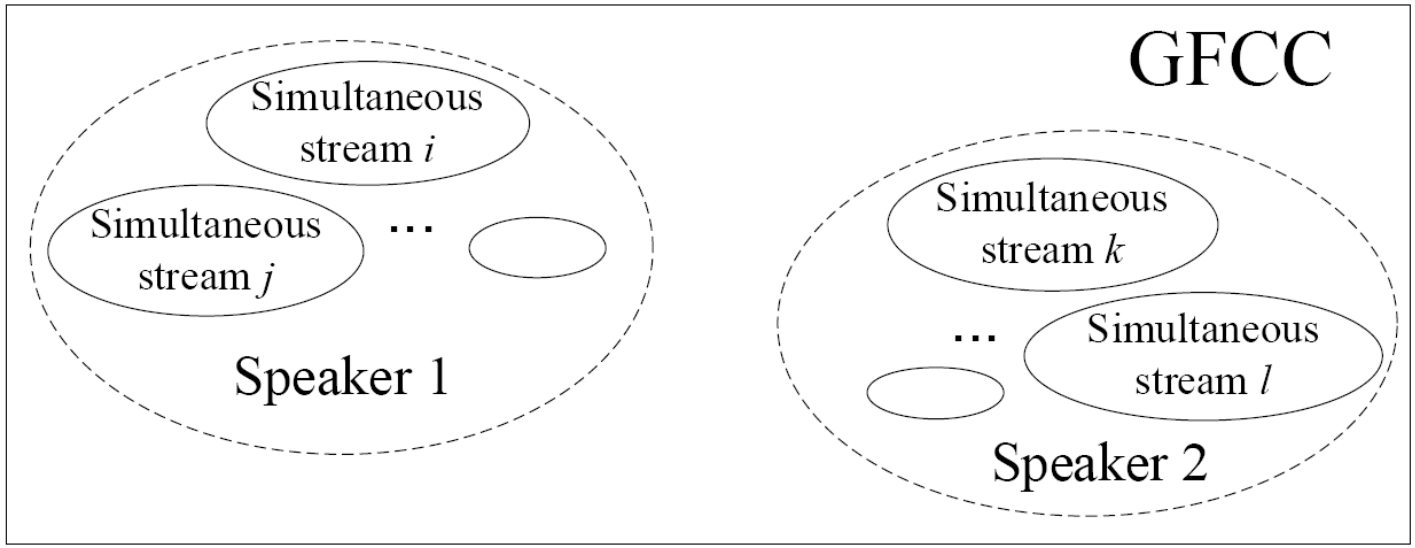


Background

- **Simultaneous streams from a cochannel (two-talker) mixture are produced using the tandem algorithm (Hu & Wang'10)**
- **Challenges**
  - Simultaneous streams consist of spectrally incomplete frames
  - Simultaneous streams are often short in duration
  - Unvoiced speech lacks harmonic structure and is relatively weak
- **Existing work uses model-based methods**
  - Typically use pretrained speaker models
  - Only work when occurring speakers are correctly identified as the corresponding trained models

Proposed idea

- **Sequential grouping based on unsupervised clustering**
  - Simultaneous streams are clustered into two groups that exhibit the largest speaker difference
  - Use gammatone frequency cepstral coefficients (GFCC) as features
    - GFCC converts gammatone filter responses using discrete cosine transform (Shao'07)



Objective function in clustering

- **Given a hypothesized grouping *g*, simultaneous streams are divided into two groups, and we calculate**

$$\mathbf{S}_{W,\mathbf{g}} = \sum_{l=1}^{2} \sum_{\mathbf{y} \in C_l} (\mathbf{y} - \mathbf{m}_l)(\mathbf{y} - \mathbf{m}_l)^t \qquad \mathbf{S}_{B,\mathbf{g}} = \sum_{l=1}^{2} N_l (\mathbf{m}_l - \mathbf{m})(\mathbf{m}_l - \mathbf{m})^t$$

$\mathbf{g}$: a $1 \times M$ binary label vector for M simultaneous streams

$\mathbf{y}$: a GFCC feature vector; $\mathbf{m}_l$: mean of lth group; $\mathbf{m}$: global mean

$\mathbf{S}_{W,\mathbf{g}}$: Within-group scatter matrix of GFCCs given $\mathbf{g}$

$\mathbf{S}_{B,\mathbf{g}}$: Between-group scatter matrix of GFCCs given $\mathbf{g}$

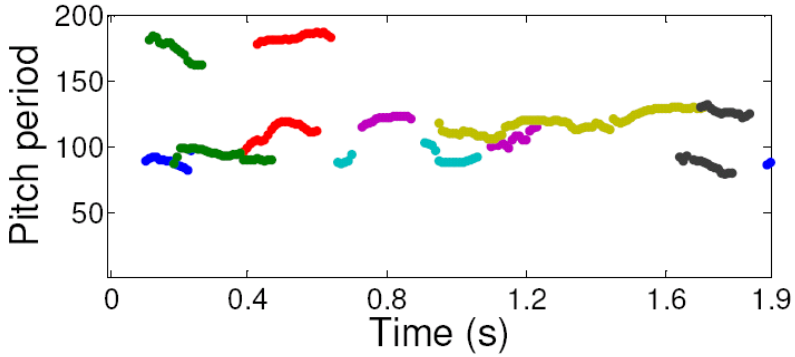
- **The trace of the ratio of between-group and within-group scatter matrices is used to measure the group difference**

$$O(\mathbf{g}) = \mathrm{tr}(\mathbf{S}_{W,\mathbf{g}}^{-1} \mathbf{S}_{B,\mathbf{g}})$$

  - The trace operator measures the ratios along eigenvector dimensions

Penalty term

- **Constraint: two simultaneous streams with overlapping pitch contours should not be assigned to the same speaker**



- **We penalize grouping *g* with *m* frames of pitch overlap in the same cluster by**

$$P(\mathbf{g}) = 1 \Big/ (1 + e^{a(m_{\mathbf{g}} - b)}), \quad a < 0 \text{ and } b \geq 0$$

$m_{\mathbf{g}}$: Number of within-cluster overlapping-pitch frames in **g**

a: Steepness of the penalty; b: Tolerance of overlapping

Constrained objective function

- **Adding the penalty term, the objective function becomes**

$$J(\mathbf{g}) = \lambda O(\mathbf{g}) - (1-\lambda) c P(\mathbf{g}), \quad 0 \le \lambda \le 1$$

  - c is a constant scaling $P(\mathbf{g})$ to the same value range as $O(\mathbf{g})$
  - λ is a tradeoff between the group difference and the penalty

Solution via search

- **We find the grouping solution g_s by maximizing $J(g)$, i.e.**

$$\mathbf{g}_s = \arg\max_{\mathbf{g} \in G} J(\mathbf{g})$$

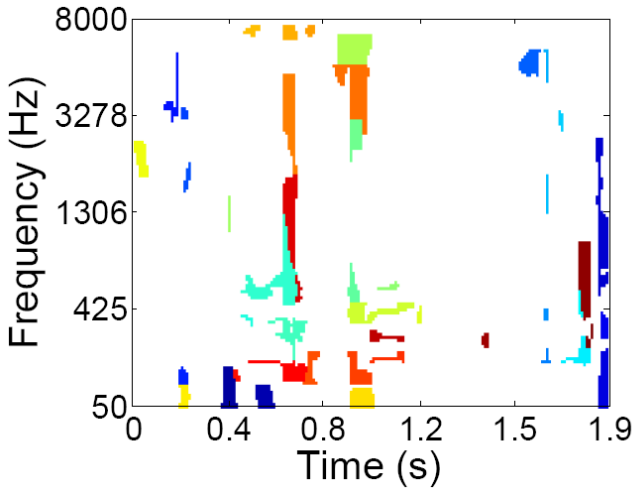
- **We use a genetic algorithm (GA) based search**
  - Each binary vector **g** is considered as a chromosome in GA
  - The objective function $J(\mathbf{g})$ is used to calculate the fitness score in GA
  - The chromosome with the highest fitness score in the final population is taken as the solution

Segregation of unvoiced speech

- **Unvoiced speech corresponds to a subset of consonants and accounts for about 20-25% of spoken English (Hu & Wang'08)**
  - Stops: /t/, /p/, /k/, **/d/**, **/b/**, and **/g/**
  - Fricatives: /s/, /f/, /ʃ/, /θ/, **/v/**, **/ʒ/**, **/z/**, **/ð/**, and **/h/**
  - Affricates: /tʃ/ and **/dʒ/**
- **Unsupervised sequential grouping of unvoiced speech is particularly difficult**
  - Noise like
  - Weak in energy

Segmentation of unvoiced speech

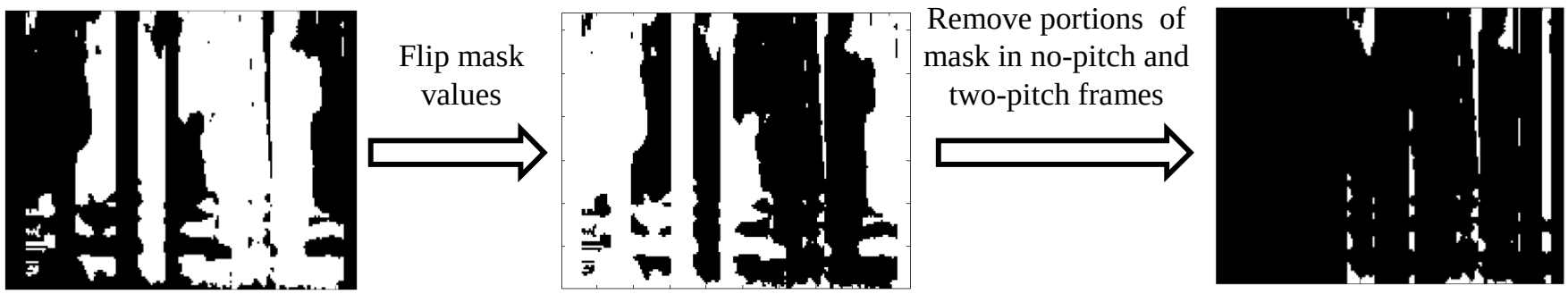
- **Speech segments are first generated using a multiscale onset/offset analysis (Hu & Wang'07)**
- **The portions of segments overlapping with voiced speech are removed**
- **Remaining parts correspond to unvoiced segments**



An example of estimated unvoiced speech segments

Grouping

- **Key idea: Label unvoiced segments by the complementary mask of a segregated voiced stream**



Segregated voiced streams of speaker *a*

The complementary mask

complementary mask in single-pitch frames, *C*

- **For each unvoiced segment, we calculate its overlapping energy with C_a and C_b to yield E_a and E_b, respectively.**
- **The segment is assigned to speaker *a* if $E_b > E_a$ and to speaker *b* otherwise**

Limitation

- **The above method deals with only unvoiced-voiced (or voiced-unvoiced) portions of the mixture but not unvoiced-unvoiced portions**
- **In the speech separation challenge (SSC) corpus (Cooke & Lee'06)**
  - Unvoiced-unvoiced portions constitute only about 10% of unvoiced speech energy in a mixture
  - Future research topic

Evaluation of voiced speech segregation

- **Test data: 100 0-dB cochannel mixtures from the SSC corpus**
  - Two types of simultaneous streams are used: estimated using the tandem algorithm and ideal
- **Evaluation metric is the SNR using the ideal binary mask as ground truth (Wang'05)**
- **Compared to a model based method (Shao & Wang'09)**
  - The method needs to know the target speaker identity and uses a background model (BM) for the other speaker

Comparisons of output SNRs (in dB) for voiced speech

Speech type	ESS			ISS		
	SG	DG	Both	SG	DG	Both
BM	3.7	6.0	4.8	8.9	11.8	10.3
Proposed	3.7	6.8	5.2	11.4	15.2	13.3
ISG	5.7	8.0	6.9	14.4	15.7	15.0

ESS: estimated simultaneous streams; **ISS**: ideal simultaneous streams; **ISG**: ideal sequential grouping

Evaluation of unvoiced speech segregation

- **We compare to a model-based method which groups unvoiced speech segments based on the detected target speaker (from the voiced part) and a background model (Shao *et al.*'10)**
  - Evaluation based on the SNR gain in unvoiced intervals
  - Our method performs comparably when using estimated simultaneous streams for voiced speech segregation
  - 3.9 dB better when using ideal simultaneous streams for sequential grouping of voiced speech

Comparisons of SNR gains (in dB) in unvoiced intervals

Speech type	ESS+UNVOICED			ISS+UNVOICED		
	SG	DG	Both	SG	DG	Both
BM	6.0	10.0	7.9	7.7	10.5	9.1
Proposed	6.0	9.7	7.8	12.1	14.0	13.0
ISG	11.6	15.4	13.5	18.0	18.1	18.0

Evaluation of overall segregation

- **Our method performs better than the model-based method by 0.7 dB and 6.3 dB in estimated and ideal cases, respectively**
- **Unvoiced speech contributes to overall segregation by 0.5 dB and 3.7 dB in the two cases**

Comparisons of overall output SNRs (in dB)

Speech type	ESS			ISS			ESS+UNVOICED			ISS+UNVOICED		
	SG	DG	Both	SG	DG	Both	SG	DG	Both	SG	DG	Both
BM	3.7	6.0	4.8	8.9	11.8	10.3	3.9	6.2	5.0	9.2	12.3	10.7
Proposed	3.7	6.8	5.2	11.4	15.2	13.3	3.9	7.5	5.7	15.1	19.0	17.0
ISG	5.7	8.0	6.9	14.4	15.7	15.0	7.9	10.6	9.2	22.9	23.1	23.0







Conclusion

- **We have proposed a novel unsupervised approach to sequential grouping in cochannel speech**
  - Clustering is used for sequential grouping of voiced speech
  - Unvoiced speech is grouped based on the complementary masks of two segregated voiced streams
- **Evaluations show that our method performs better than a previous model-based approach**