An SVM Based Classification Approach To Speech Separation

Kun Han and DeLiang Wang

Perception & Neurodynamics Laboratory
Ohio State University
May 2011

Outline of presentation

- **Introduction**
- **Feature extraction**
- **Unit labeling and segmentation**
- **Experiments**

Monaural speech separation

- **In a daily environment, target speech is often corrupted by various types of acoustic interference**
- **How to remove or attenuate background noise?**
- **Monaural speech separation**
  - One has to consider the **intrinsic properties** of target speech and interference

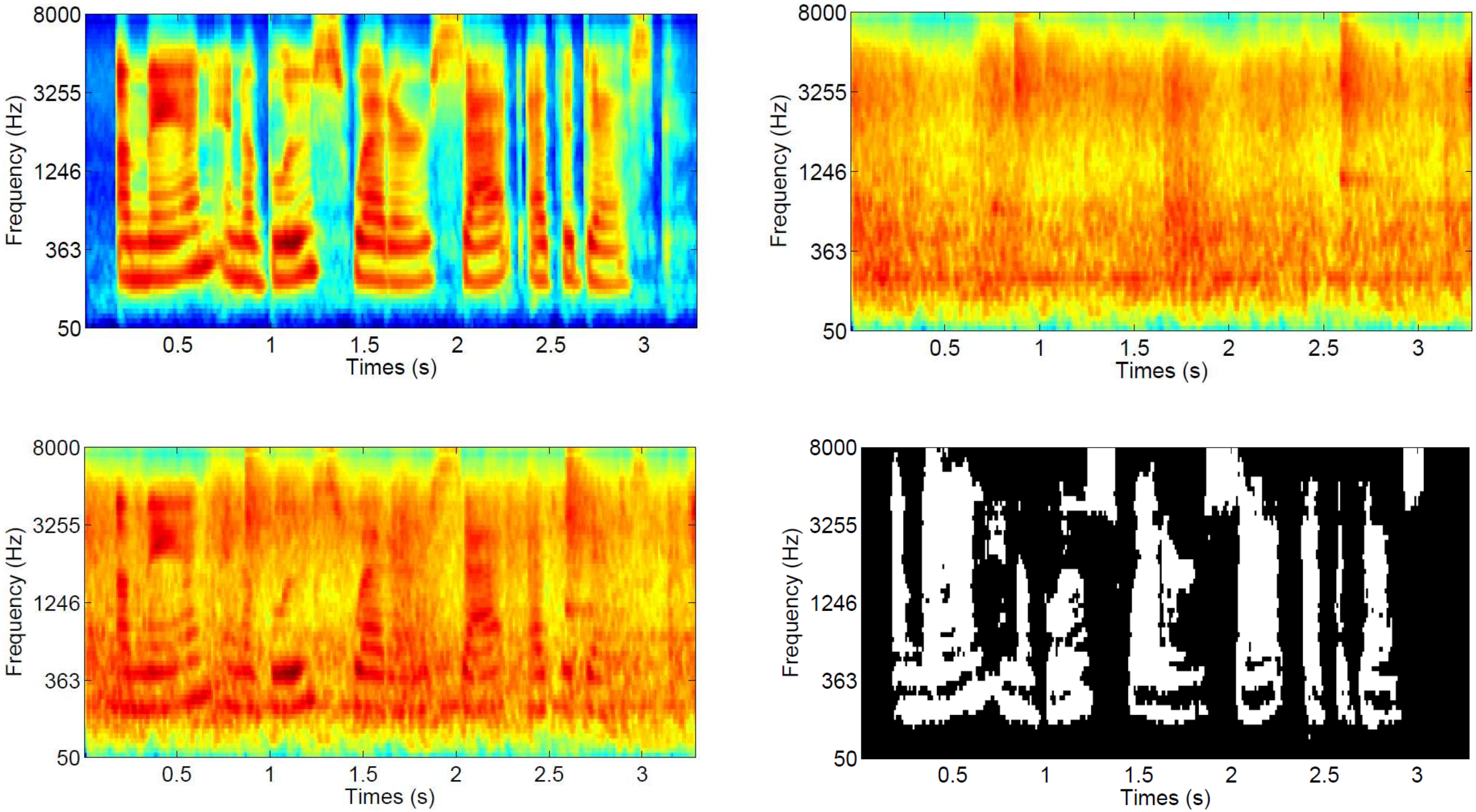
Ideal binary mask

- **A main computational goal for CASA: Ideal binary mask (IBM)**
- **The definition of the IBM (Wang'05):**

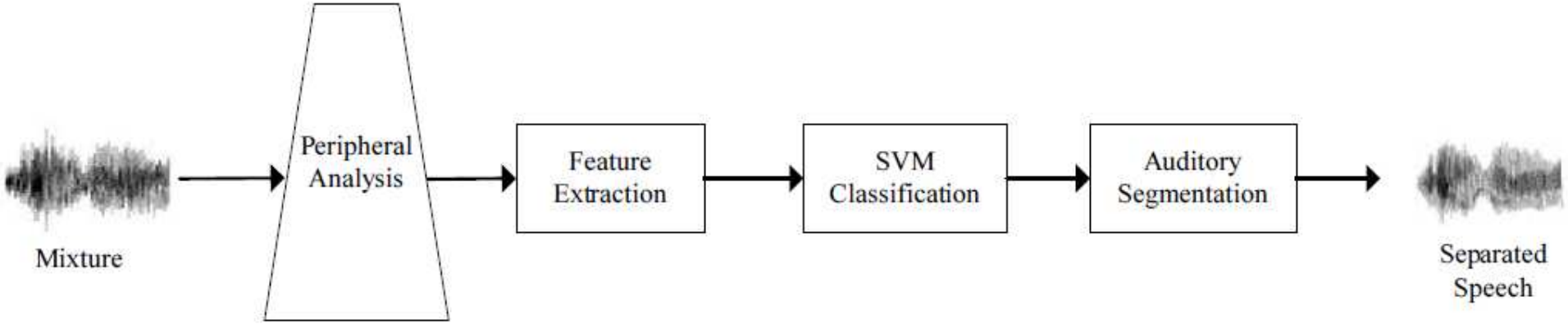
$$IBM(t, f) = \begin{cases} 1 & \text{if SNR}(t, f) > \theta \\ 0 & \text{otherwise} \end{cases}$$

  - θ: Local SNR criterion (LC)
  - **Speech separation → IBM estimation →binary classification**

Ideal binary mask illustration



System overview



- **Feature extraction**
  - Pitch-based
  - Amplitude modulation spectrum (AMS)
- **Unit labeling**
  - Support vector machine
  - Re-thresholding
- **Segmentation**

Pitch-based features

- **For each T-F unit, we compute autocorrelation $A(c, m, \tau_m)$ at pitch lag τ_m ($A(c, m, \tau_m) = 0$ for unvoiced frames)**
- **Use delta features to capture feature variations across time and frequency**

$$\Delta A^T(c,m,\tau_m) = A(c,m,\tau_m) - A(c,m-1,\tau_m)$$

$$\Delta A^C(c,m,\tau_m) = A(c,m,\tau_m) - A(c-1,m,\tau_m)$$

- **Also compute envelope autocorrelation $A_E(c, m, \tau_m)$ and its delta features**

$$\mathbf{x}_A(c,m) = \begin{pmatrix} A(c,m,\tau_m) \\ A_E(c,m,\tau_m) \\ \Delta A^T(c,m,\tau_m) \\ \Delta A_E^T(c,m,\tau_m) \\ \Delta A^C(c,m,\tau_m) \\ \Delta A_E^C(c,m,\tau_m) \end{pmatrix}$$

AMS features

- **For each T-F unit, we extract a 15-dimensional AMS feature (Kim et al.'09)**
  - $[M_1(c, m), \ldots, M_{15}(c, m)]$
- **Similarly, we calculate delta features**

$$\mathbf{x}_M(c,m) = \begin{pmatrix} M_1(c,m) \\ \cdots \\ M_{15}(c,m) \\ \Delta M_1^T(c,m) \\ \cdots \\ \Delta M_{15}^T(c,m) \\ \Delta M_1^C(c,m) \\ \cdots \\ \Delta M_{15}^C(c,m) \end{pmatrix}$$

Support vector machine (SVM)

- **Input: Feature vectors from T-F units**
  - pitch-based + AMS
- **Discriminant function of SVM**

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$$

$$y(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$$

- **Train an SVM for each channel (64 channels)**
  - Gaussian kernel
  - Parameters: C and σ are chosen using 5-fold cross validation

Measurement: HIT-FA

	IBM	Estimated IBM
Reject	0	0
False alarm	0	1
Hit	1	1
Miss	1	0

- HIT rate = Hit / (Hit + miss)
- FA rate = false alarm / (Reject + false alarm)
- HIT-FA is highly correlated to speech intelligibility (Kim et al.’09)

Re-thresholding

- **Problems:**
  - Classification accuracy or HIT-FA?

- **Re-thresholding:**
  - Instead of 0, we choose a new threshold to maximize HIT-FA in each channel

$$y(\mathbf{x}) = \begin{cases} 1, \text{ if } f(\mathbf{x}) > \theta_c \\ 0, \text{ otherwise} \end{cases}$$

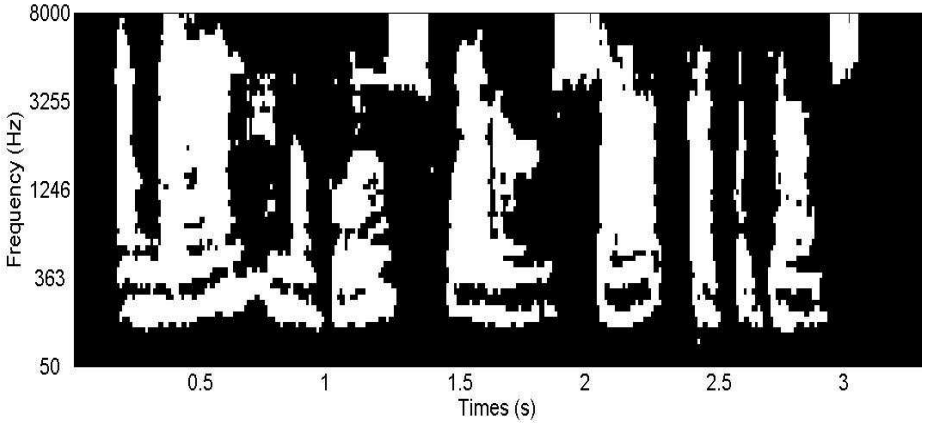
  - θ_c is chosen from a small validation set

Auditory segmentation

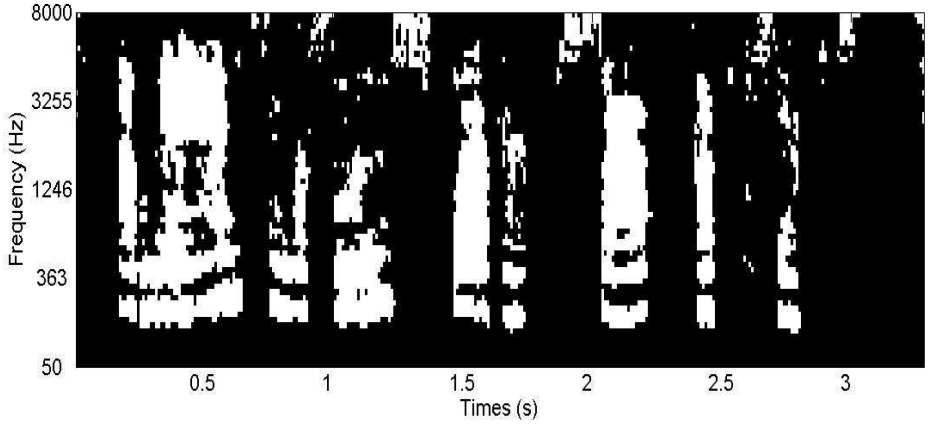
- **Voiced frames**
  - Cross-channel correlation and envelope cross-channel correlation
- **Unvoiced frames**
  - Onset/offset analysis

Estimated IBM

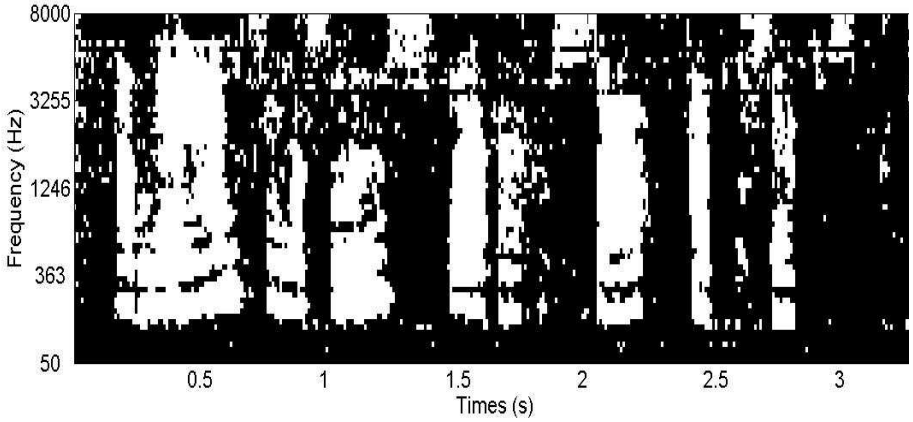
IBM



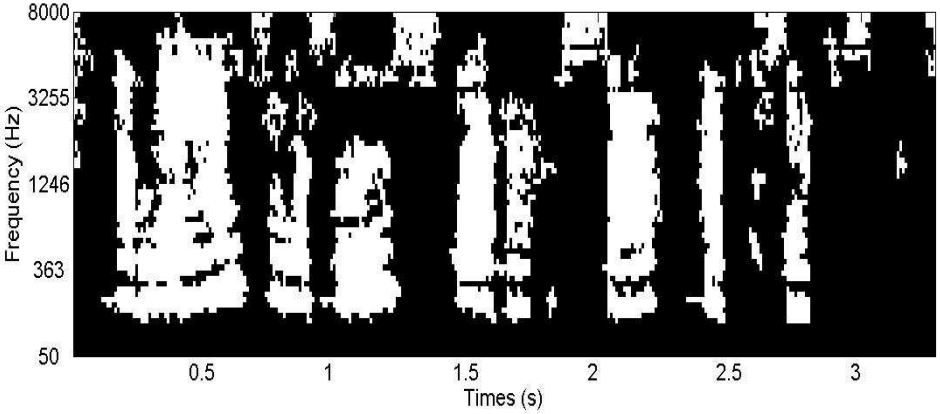
SVM labeling



Re-thresholding



Segmentation



Evaluation

- **Training**
  - 100 utterances from IEEE corpus (female)
  - Noise: Speech-shaped, factory, babble
  - Ground-truth pitch extracted from target speech
  - SNR = -5, 0, 5 dB
- **Test**
  - 60 utterances from IEEE corpus
  - Noise: Speech-shaped, factory, and babble, plus white and cocktail-party
  - Estimated pitch (Jin et al.'11)
  - SNR = -5, 0 dB

Classification results

Table 1. Classification results for different noises

		Speech-shaped		Factory		Babble	
		-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
Proposed	**HIT**	60.14%	69.89%	60.02%	70.52%	61.43%	69.00%
	FA	4.10%	3.89%	8.60%	7.09%	17.58%	16.12%
	HIT-FA	56.04%	66.00%	51.42%	63.43%	43.85%	52.88%
	Accuracy	90.33%	89.60%	86.09%	87.02%	77.52%	78.63%
Kim *et al.*	**HIT**	59.74%	61.02%	57.39%	60.38%	53.85%	56.30%
	FA	20.70%	16.20%	26.71%	22.43%	27.18%	24.60%
	HIT-FA	39.04%	44.82%	30.68%	37.95%	26.67%	31.71%
	Accuracy	76.25%	78.15%	70.60%	73.05%	68.40%	68.86%

Table 2. Classification results for new noises

		White		Cocktail-party	
		-5 dB	0 dB	-5 dB	0 dB
Proposed	**HIT**	69.44%	72.55%	54.31%	66.29%
	FA	7.25%	8.32%	7.02%	6.27%
	HIT-FA	62.19%	64.23%	47.29%	60.02%
	Accuracy	88.81%	87.00%	83.34%	84.03%
Kim *et al.*	**HIT**	48.32%	56.40%	55.43%	58.54%
	FA	25.80%	25.61%	29.13%	24.36%
	HIT-FA	22.52%	30.78%	26.31%	34.17%
	Accuracy	69.83%	69.99%	67.03%	69.60%

- Compared to Kim et al.'09 system (AMS+GMM) which improves speech intelligibility in listening tests, our system achieves higher HIT-FA rates
  - Particularly for unseen noises

Classifier comparison

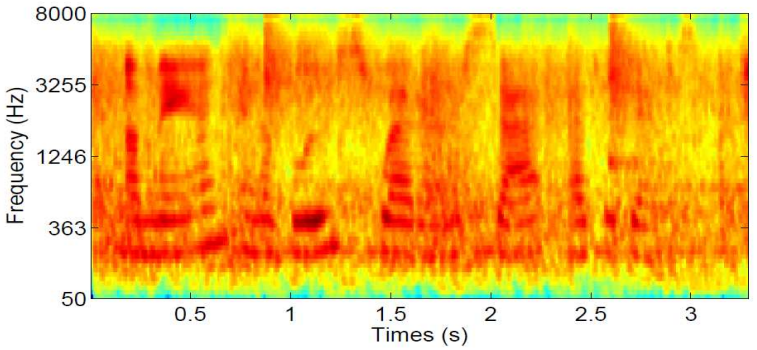
Table 3. Classification results with AMS features

		Speech-shaped		Factory		Babble	
		-5 dB	**0 dB**	**-5 dB**	**0 dB**	**-5 dB**	**0 dB**
SVM	**HIT**	77.51%	82.87%	74.26%	82.25%	80.84%	83.08%
	FA	8.43%	10.10%	12.89%	13.89%	15.80%	16.60%
	HIT-FA	69.08%	72.77%	61.37%	68.35%	65.04%	66.48%
GMM	**HIT**	80.84%	79.64%	81.91%	81.35%	81.36%	78.40%
	FA	13.27%	14.70%	24.01%	21.89%	16.57%	16.44%
	HIT-FA	67.57%	64.94%	57.90%	59.46%	64.79%	61.96%

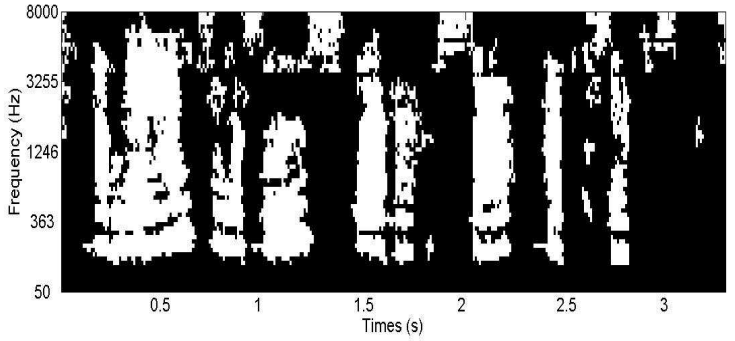
- 25-channel mel-scale filterbank
- These improvements demonstrate the advantage of SVM over GMM classifier

Demo

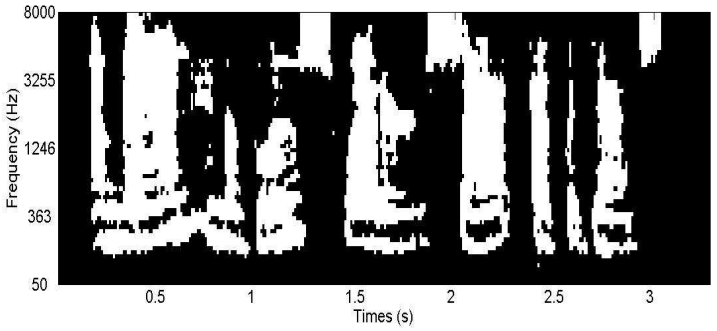
- **Female Speech + Factory Noise (0 dB)**
  - Noisy speech



  - Proposed



  - IBM



Summary

- **We treat speech separation as binary classification**
- **Use SVM to classify T-F units using pitch-based and AMS features**
- **Based on comparisons, we predict that our separation results will lead to significantly improved speech intelligibility in noisy conditions for human listeners**
  - Future research will test this prediction

- **Thank you!**