Towards Automatic Speech Recognition In Adverse Environments

D. Dimitriadis, A. Katsamanis, P. Maragos, G. Papandreou and V. Pitsikalis

NTUA, School of ECE, http://cvsp.cs.ntua.gr

Computer Vision, Speech Communication and Signal Processing Research Group
Outline

- Nonlinear Speech Processing: Background and Recent work
  - Modulations
  - Fractals
- Audio-Visual Processing
- Adaptation
- Applications to Robust ASR
Physics of speech airflow → **LINEAR ACOUSTICS APPROXIMATION** → Linear models of speech production
Evidence for Speech Modulations

- separated & unstable airflow
- vortices
- oscillators with time-varying elements
- energy pulses (Teager)
Speech Modulation Model

(Maragos, Kaiser & Quatieri 1991)

- One Single Resonance as damped AM–FM:

\[ S(t) = A(t)e^{-\alpha t} \cos \left( \omega_c t + \int_0^t q(\tau)d\tau + \theta \right) \]

Inst. Frequency: \( \dot{\omega}(t) = \omega(t)q(t) \frac{dt}{dt} \)

- If due to 2nd-order LTI system

\( A(t) = \text{constant}, \quad \omega(t) = \omega_c \)

- Speech Signal as multi-component AM-FM:

\[ \text{Speech}(t) \approx \sum_k a_k(t) \cos(\phi_k(t)) \]
Energy Tracking in Oscillators

- **harmonic oscillator**

![Harmonic oscillator diagram](image)

- **motion equation**

\[ m\ddot{x} + kx = 0 \]

- **response**

\[ x(t) = A \cos(\omega t + \theta), \quad \omega^2 = k/m \]

- **energy**

\[ E = \frac{1}{2} m\dot{x}^2 + \frac{1}{2} kx^2 = \frac{m}{2} (A^2 \omega^2) = \text{constant} \]

- **energy tracking**

\[ \Psi(x) = (\dot{x})^2 - x\ddot{x} = A^2 \omega^2 = \frac{E}{(m/2)} \]
Continuous-time signals $x(t)$:

property:

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$$

Discrete-time signals $x(n)$:

property:

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1)$$
Energy Separation Algorithm (ESA)

(Maragos, Kaiser & Quatieri 1991)

- **Cosine:**
  \[ x(t) = A \cos (\omega_c(t) + \theta) \]

  \[ \Psi[x(t)] = A^2 \omega_c^2 \]

- **AM-FM signal:**
  \[ x(t) = a(t) \cos (\int_0^t \omega(\tau) d\tau) \]

  \[ a(t), \omega(t) \text{ do not vary too fast or too much w.r.t. } \omega_c \]

\[
\begin{align*}
\Psi[\dot{x}(t)] &= A^2 \omega_c^4 \\
\frac{-[\psi]}{\sqrt{-[\Omega]}} \{ |a(t)| \} &\quad \sqrt{-[\Omega]} \{ T(t) \}
\end{align*}
\]
ESA Applied to Speech Resonance

[Graphs and diagrams related to speech signal, spectrum, and bandpass speech are shown.]
Multiband Demodulation and F/B Tracking

(Potamianos & Maragos 1996)
Frequency and Bandwidth Estimates

- Center Frequency Estimates:

\[ F_{\text{fft}} \left( \frac{1}{T} \right) \]

\[ F_w = \frac{T \int f(0) \, dt}{T \int a(t)^2 \, dt} \]

- Bandwidth Estimates:

\[ \frac{B_{\text{fft}}^2}{T} \left( \frac{1}{T} \right) \]

\[ B_w^2 = \frac{\int \omega^2 \text{M}^2 \, d\omega}{T \int a(t)^2 \, dt} \]
Modulation Acoustic Features

AM-FM Modulation Features:
- Mean Inst. Ampl.  \( \text{IA-Mean} \)
- Mean Inst. Freq.  \( \text{IF-Mean} \)
- Freq. Mod. Percent.  \( \text{FMP} \)

Energy Features:
- Teager Energy Cepstrum Coeff.  \( \text{TECC} \)
Auditory Filter-banks

- Dense asymmetrical filters estimate activity in each frequency band
- Equivalent Rectangular Bandwidth (ERB) used to quantify bandwidth of filters
- Gammatone filters used to approximate auditory filters
- Bi-parametric filter-bank design
  - Parameter 1: Number of filters
  - Parameter 2: Filter bandwidth as percent of ERB
Teager Energy Cepstrum Coefficients (TECCs)

- Use a Gammatone filterbank (with 20-40 filters) to bandpass the speech signal. The filter spacing is linear in bark scale.

- Estimate the logarithm of short-time average of the Teager-Kaiser energy operator for each band-passed signal. The short-time averaging window duration and window shift are the same as for the standard MFCC front-end.

- Estimate the cepstrum coefficients of the short-time average Teager Energy using the discrete cosine transform (DCT).

- Truncate the TE cepstrum coefficients to keep the first 13 coefficients (including the zeroth coefficient C0), similarly to the standard MFCC front-end.
Speech Recognition Experiments

- **TIMIT**
  - Phone recognition task
  - Mono-phone ASR models—no grammar

- **TIMIT + Noise**
  - Phone recognition task
  - Noise added artificially from NOISEX dB

- **AURORA 3 Spanish Task**
  - Connected digit recognition task
  - Whole digit ASR models
Results: TIMIT+Noise

Up to +106%
Results: Aurora 3 (HTK)

Up to +62%
## ASR Results

### TIMIT-Based Speech Databases

(Correct Phone Accuracies (%))

<table>
<thead>
<tr>
<th>Features</th>
<th>TIMIT</th>
<th>NTIMIT</th>
<th>TIMIT+ Babble</th>
<th>TIMIT+ White</th>
<th>TIMIT+ Pink</th>
<th>TIMIT+ Car</th>
<th>Av. Rel. Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC*</td>
<td>58.40</td>
<td>42.42</td>
<td>27.71</td>
<td>17.72</td>
<td>18.60</td>
<td>52.75</td>
<td>-</td>
</tr>
<tr>
<td>TEner. CC</td>
<td>58.89</td>
<td>42.40</td>
<td>41.61</td>
<td>34.74</td>
<td>38.40</td>
<td>54.35</td>
<td>24.26</td>
</tr>
<tr>
<td>MFCC*+ IA-Mean</td>
<td>59.61</td>
<td>43.53</td>
<td>39.25</td>
<td>26.03</td>
<td>31.05</td>
<td>56.50</td>
<td>17.62</td>
</tr>
<tr>
<td>MFCC*+ IF-Mean</td>
<td>59.34</td>
<td>43.70</td>
<td>36.87</td>
<td>25.38</td>
<td>30.92</td>
<td>55.30</td>
<td>15.58</td>
</tr>
<tr>
<td>MFCC*+ FMP</td>
<td>59.92</td>
<td>43.69</td>
<td>38.60</td>
<td>26.15</td>
<td>32.84</td>
<td>55.97</td>
<td>18.17</td>
</tr>
</tbody>
</table>

* MFCC+C₀+D+DD,  # states=3,  # mixtures=16

---

**NTUA**

Cost Meeting, Heraklio Greece Sep. 2005
# Experimental Results IIa (HTK)

## Aurora3 (Spanish Task)

(Correct Word Accuracies (%))

<table>
<thead>
<tr>
<th>Features</th>
<th>WM</th>
<th>MM</th>
<th>HM</th>
<th>Average</th>
<th>Av. Rel. Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aurora Front-End (WI007)</strong></td>
<td>92.94</td>
<td>80.31</td>
<td>51.55</td>
<td>74.93</td>
<td>-</td>
</tr>
<tr>
<td>MFCC*</td>
<td>93.68</td>
<td>92.73</td>
<td>65.18</td>
<td>83.86</td>
<td><strong>+35.62 %</strong></td>
</tr>
<tr>
<td>TEnerCC+log (Ener) +CMS</td>
<td>93.64</td>
<td>91.61</td>
<td>86.85</td>
<td>90.70</td>
<td><strong>62.90 %</strong></td>
</tr>
<tr>
<td>MFCC*+IA-Mean</td>
<td>94.05</td>
<td>92.22</td>
<td>77.70</td>
<td>87.99</td>
<td><strong>52.09 %</strong></td>
</tr>
<tr>
<td>MFCC*+IF-Mean</td>
<td>90.71</td>
<td>89.52</td>
<td>72.36</td>
<td>84.20</td>
<td><strong>36.98 %</strong></td>
</tr>
<tr>
<td>MFCC*+FMP</td>
<td>94.41</td>
<td>92.46</td>
<td>82.73</td>
<td>89.87</td>
<td><strong>59.59 %</strong></td>
</tr>
</tbody>
</table>

* MFCC+log(Ener)+D+DD+CMS, # states=14, # mixtures=14

---

*NTUA*

Cost Meeting, Heraklio Greece Sep. 2005
Future Work on Modulation Features

- Refinements w.r.t. AM-FM Features
- Fusion w. Other Features
Fractal Features

speech signal $\rightarrow$ Embedding $\rightarrow$ N-d Signal $\rightarrow$ Local SVD $\rightarrow$ N-d Cleaned

Geometrical Filtering $\rightarrow$ MFD

Multiscale Fractal Dimension (6)

Filtered Dynamics - Correlation Dimension (8)

Neighborhood Distance Reduction

Noisy Embedding $\rightarrow$ Filtered Embedding

Median Neighborhood Distance

NTUA

Cost Meeting, Heraklio Greece Sep. 2005
Reconstructed Attractors
Fractal Features: Correlation Dimension

Correlation Dimension:
\[ D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C_N(r)}{\log r} \]

Correlation sum:
\[ C_N(r;x) = \sum_{ij} \left[ f(2r/\|x_i-x_j\|) \right] \]

Histogram of \( D_C \) component for selected phoneme classes (stops, vowels, fricatives)

- \( N \): # of points, \( r \): scale
- \( x \): set points
- \( f \): heavyside function
Morphological Measurement of Fractal Dimension

- **Minkowski cover of curve**

- **Fractal (Minkowski-Bouligand) dimension**

\[ D \in [1, 2] \]

\[ A_B(r) = \text{Length}(G) \]

\[ \frac{A_B(r)}{2r} \]

- **Least-Squares line fit to data**

\[ \left( \log, \frac{1}{\text{Area}(r)} \right)^2 \]
Multiscale Speech Fractal Dimension

- short-time speech signal
  \[ S(t), 0 \leq t \leq T \]
- signal graph
  \[ G_{ts}(\theta_{0}, (\theta_{0}))^{2} \]
  \[ \Rightarrow \]
- fractal constant power law
  \[ \text{area}(GB(C \cup J))^{22D} \text{, as } 0 \]

- variable power law
  \[ \text{area}(GB(C \cup J))^{22D(j)} \]

- multiscale fractal “dimension” (speech fractogram):
  \[ MFD(t, \varepsilon) = D(\varepsilon) \]
  of short-time speech segment around time \( t \)
Multiscale Fractal Dimension

SPEECH SIGNAL: /f/  

FRACTAL DIMENSION of /f/

SPEECH SIGNAL: /v/  

FRACTAL DIMENSION of /v/

SPEECH SIGNAL: /iy/  

FRACTAL DIMENSION of /iy/

/lf/  /lv/  /liy/
Mean MFD for /sh/, /zh/, /uh/, /t/, /d/
### Word Percent Correct For the E-set Recognition Task

(ISOLET Database, 5-Mixture Gaussians per HMM State)

<table>
<thead>
<tr>
<th>Features</th>
<th>Models</th>
<th>5-mixture Gaussians</th>
<th>10-mixture Gaussians</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {E_C, E \cap \infty } )</td>
<td>81.2%</td>
<td>85.6%</td>
<td>88.6%</td>
</tr>
<tr>
<td>+ ( {D_1, \Delta D_1 } )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( {E_C, E \cap \infty } )</td>
<td>83.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ ( {D_{14-61-16} } )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( {E_C, E \cap \infty } )</td>
<td>84.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Task: Speaker Independent Recognition of Digit Sequences

TI - Digits at 8kHz

Training (8440 Utterances per scenario, 55M/55F)
- Clean (8kHz, G712)
- Multi-Condition (8kHz, G712)
  - 4 Noises (artificial): subway, babble, car, exhibition
  - 5 SNRs: 5, 10, 15, 20dB, clean

Testing, artificially added noise
- 7 SNRs: [-5, 0, 5, 10, 15, 20dB, clean]
- A: noises as in multi-cond train., G712 (28028 Utters)
- B: restaurant, street, airport, train station, G712 (28028 Utters)
- C: subway, street (MIRS) (14014 Utters)
Results: Aurora 2

Graph showing word accuracy (%) for different SNR levels (Clean, 20dB, 15dB, 10dB, 5dB, 0dB, Average) with and without FDCD. The graph indicates an improvement up to +40% with FDCD.
Results: Aurora 2

Up to +27%

Word Accuracy

SNR

clean  20 dB  15 dB  10 dB  5 dB  0 dB  Ave.

Baseline  MFD

Cost Meeting, Heraklio Greece Sep. 2005
Results: Aurora 2

Up to +61%
Future Directions on Fractal Features

- Refine Fractal Feature Extraction.
- Application to Aurora 3.
- Fusion with other features.
Visual Front-End

- **Aim:**
  - Extract low-dimensional visual speech feature vector from video

- **Visual front-end modules:**
  - Speaker's face detection
  - ROI tracking
  - Facial Model Fitting
  - Visual feature extraction

- **Challenges:**
  - Very high dimensional signal - which features are proper?
  - Robustness
  - Computational Efficiency
Face Modeling

- A well studied problem in Computer Vision:
  - Active Appearance Models, Morphable Models, Active Blobs
- Both *Shape & Appearance* can enhance lipreading
- The shape and appearance of human faces “live” in low dimensional manifolds
Image Fitting Example

step 2

step 6

step 10

step 14

step 18
Example: Face Interpretation Using AAM

- Original video
- Shape track superimposed on original video
- Reconstructed face

This is what the visual-only speech recognizer “sees”!

- Generative models like AAM allow us to evaluate the output of the visual front-end
Evaluation on the CUAVE Database
Audio-Visual ASR: Database

- Subset of CUAVE database used:
  - 36 speakers (30 training, 6 testing)
  - 5 sequences of 10 connected digits per speaker
  - Training set: 1500 digits (30x5x10)
  - Test set: 300 digits (6x5x10)

- CUAVE database also contains more complex data sets:
  speaker moving around, speaker shows profile, continuous digits, two speakers (to be used in future evaluations)

- CUAVE was kindly provided by the Clemson University
Recognition Results (Word Accuracy)

- Data
  - Training: ~500 digits (29 speakers)
  - Testing: ~100 digits (4 speakers)

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Visual</th>
<th>Audiovisual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>99%</td>
<td>46%</td>
<td>85%</td>
</tr>
<tr>
<td>Recognition</td>
<td>98%</td>
<td>26%</td>
<td>78%</td>
</tr>
</tbody>
</table>
Future Work

- Visual Front-end
  - Better trained AAM
  - Temporal tracking

- Feature fusion
  - Experimentation with alternative DBN architectures
  - Automatic stream weight determination

- Integration with non-linear acoustic features

- Experiments on other audio-visual databases

- Systematic evaluation of visual features
User Robustness, Speaker Adaptation

- **VTLN**
  - Platform: HTK
  - Database: AURORA 4
  - Fs = 8 kHz
  - Scenarios: Training, Testing
  - Comparison with MLLR

- Collection of non-Native Speech Data
  - 10 Speakers
  - 100 Utterances/Speaker
Vocal Tract Length Normalization

Implementation: HTK

- Warping Factor Estimation
  - Maximum Likelihood (ML) criterion

- Frequency Warping

Figures from Hain99, Lee96

Cost Meeting, Heraklio Greece Sep. 2005
Training
- AURORA 4 Baseline Setup
- Clean (SIC), Multi-Condition (SIM), Noisy (SIN)

Testing
- Estimate warping factor using adaptation utterances (Supervised VTLN)
  - Per speaker warping factor (1, 2, 10, 20 Utterances)
- 2-pass Decoding
  - 1\textsuperscript{st} pass
    - Get a hypothetical transcription
  - Alignment and ML to estimate per utterance warping factor
  - 2\textsuperscript{nd} pass
    - Decode properly normalized utterance
VTLN Results, Clean Training

Word Error Rate (%)

Test Noise

Clean | Car | Tr. Station

SIC | Supervised VTLN | MLLR - 2 | MLLR-2O
VTLN Results, Multi-Condition Training

Test Noise

Word Error Rate (%)

Clean
Car
Tr. Station

SIM
Supervised VTLN
MLLR - 2

NTUA

Cost Meeting, Heraklio Greece Sep. 2005
VTLN Results, Noisy Training

Word Error Rate (%)

Test Noise

Clean  Car  Tr. Station

SIN  Supervised VTLN  MLLR - 2
Future Directions for Speaker Normalization

- Estimate warping transforms at signal level
  - Exploit instantaneous amplitude or frequency signals to estimate the warping parameters, Normalize the signal

- Effective integration with model-based adaptation techniques (*collaboration with TSI*)
Conclusions

- Results on Modulation (AM-FM and TECC) Features
- Results on Fractal (MFD and FDCD) Features
- Results on AV-ASR
- Results on Adaptation

Our Publications in  http://cvsp.cs.ntua.gr