Exploitation of Phase and Vocal Excitation Modulation Features for Robust Speaker Recognition

Ning Wang

12 September, 2011

Outline

Outline

1. Introduction
2. Exploration of excitation modulation features
3. Phase information derivation
4. Performance evaluation of modulation features
5. Extraction of robust speaker features
6. Summary

Speaker recognition

- Aim

  Speaker recognition (SR) refers to the process of automatically determining or verifying the identity of a person based on his or her voice.

  Including: Speaker identification (SID), speaker verification (SV).

- Application
  - Sole medium for identity authentication, *e.g.*, telephony banking, ...
  - One modality in multi-modal biometric system, ...

Challenges and difficulties

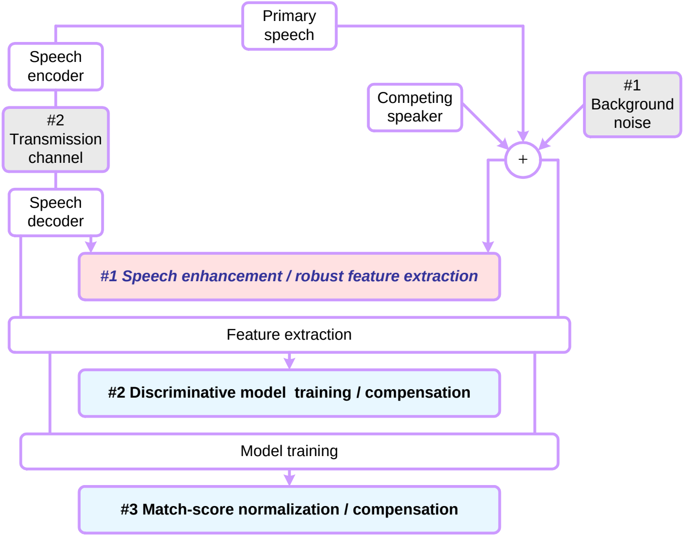
The SR system in practical applications has to handle

- Sparsity of available data set.
- Inconsistence of data quality.
- Variabilities in application scenarios.
  - Background noise.
  - Transmission channels.
  - Emotional, healthy states of speakers.

  ...

...

Robust speaker recognition



Speech feature overview

- Cepstral coefficients
  - Conventional one: Mel-frequency cepstral coefficients (MFCCs).
  - Most commonly used in both speech recognition and speaker recognition systems.
  - Primarily characterize the spectral envelope of a quasi-stationary speech segment.

Speech feature overview

- Exploration of alternative properties
  e.g.,
  - Vocal source (Prasanna, 2006; Zheng, 2007; Gudnason, 2008).
  - Phase information (Ambikairajah, 2007; Grimaldi, 2008).
  - Prosodic characteristics (Atal, 1972; Shriberg, 2005; Adami, 2007).
  - High-level features (Doddington, 2001; Andrews, 2002; Leung, 2006).
- Employment of other parameterization methods
  e.g.,
  - Wavelet transform (Zheng, 2007).
  - Time-frequency principal component analysis (Magrin-Chagnolleau, 2002).
  - Modulation analysis (Dimitriadis, 2005; Thiruvaran, 2008).

Modulation properties in speech

- Formant AM-FM modeling
  - Formant frequency and bandwidth tracking (Potamianos, 1996).
  - Modulation features for robust SR
    - Average log-envelope (Wang, 2003).
    - Dynamic spectral subband centroid (Chen, 2004).
    - Amplitude weighted instantaneous frequency (Dimitriadis, 2005).
    - FM features (Ambikairajah, 2007).
- Sinusoidal representation
  - Pitch tracking (Mcaulay, 1990; Stylianou, 1996).
  - Vocoder (Mcaulay, 1986; Potamianos, 1999).

AM-FM signal modeling

- An AM-FM signal

carrier frequency

arbitrary phase offset

$$x(n) = A(n) \times \cos[\Theta(n)] = A(n) \times \cos\left[\Omega_c n + \Omega_m \sum_{r=1}^{n} q(r) + \phi\right]$$

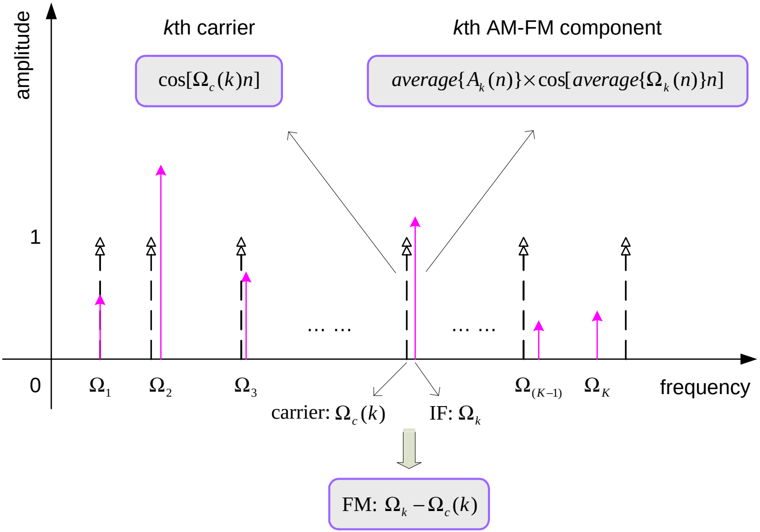
time - varying amplitude

frequency deviation constant

- Time-varying angular frequency:

$$\Omega(n) = \Theta(n) - \Theta(n-1) = \Omega_c + \Omega_m q(n) = \frac{2\pi}{fs} fc + \Omega_m q(n) \longrightarrow \text{FM component}$$

AM-FM signal representation



Motivation of work

- Modulation phenomena in vocal excitation carries speaker-distinctive characteristics.
- Phase related information in speech is important but absent from the magnitude based features.
- Complementary information source helps in improving the performance of MFCCs predominant SR systems.
- Speaker discrimination in practical applications desires robust parameters.

Research focus

- Exploration: how are the speaker-distinctive properties carried by the modulation parameters?
- Investigation: how can we parameterize the modulation parameters into speaker features?
- Evaluation: how complementary are the parameterized features with MFCC features in SR experiments?
- Refinement: how to enhance robustness of modulation features under various application scenarios?

Outline

1. Introduction
2. Exploration of excitation modulation features
3. Phase information derivation
4. Performance evaluation of modulation features
5. Extraction of robust speaker features
6. Summary

Excitation signal modeling for SR

For excitation signals,

- spectral analysis mostly concerns pitch-periodicity or harmonic structure properties.
- signal decomposition provides spectro-temporal parameters.
  e.g., in sinusoidal modeling, harmonic plus noise modeling, AM-FM modeling, etc.

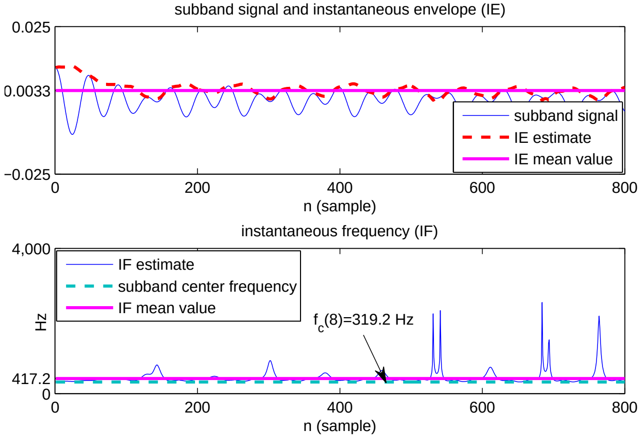
Facilitating extraction of speaker-specific characteristics, we need to

1. select proper source model to use.
2. parameterize model parameters into feature vectors.
3. inspect feature vectors in expressing typical vocal properties.

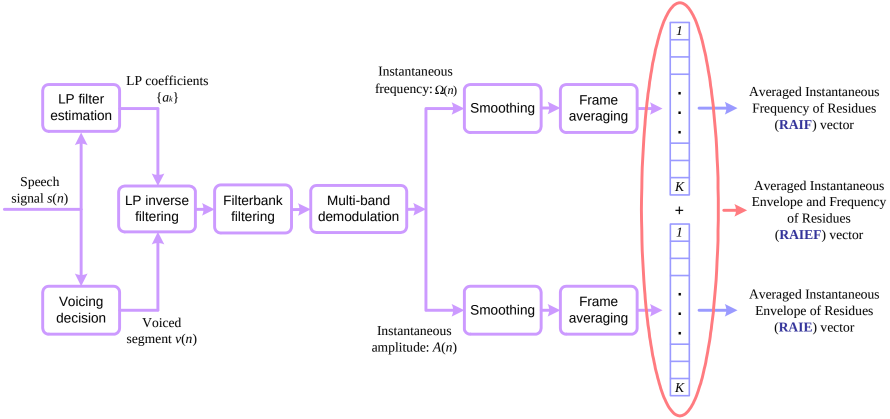
AM-FM excitation representation

Multicomponent AM-FM excitation signal modeling:

- Represents inclusive *monocomponent* signals in terms of time-varying envelope (amplitude) and frequency: $A(n)cos\left[\Theta(n)\right]$.



Excitation modulation feature derivation



- K: Subband number (center frequency in kth subband is $f_c(k)$).
- RAIEF: Combination of RAIE and RAIF vectors.

Excitation modulation features

- Speech production:

  Air flow, vocal fold open and close, vibration activity.

- Vocal excitation properties concerned:
  - Pitch period (F0) and harmonics;
  - Pitch epoch shape;
  - Details embedded between adjacent pitch epochs.

Feature analysis

- Artificially generated pulse trains: $e_1(n)$, $e_2(n)$, $e_3(n)$ and $e_4(n)$.
  - Approximate vocal excitation signals.

	$F0\ (Hz)$	Epoch shape	Details?
$e_1(n)$	86.2	Impulse	No
$e_2(n)$	172.4	Impulse	No
$e_3(n)$	172.4	Triangular pulse	No
$e_4(n)$	172.4	Triangular pulse	Yes

  - F0 values and center frequencies in some bands specially settled.

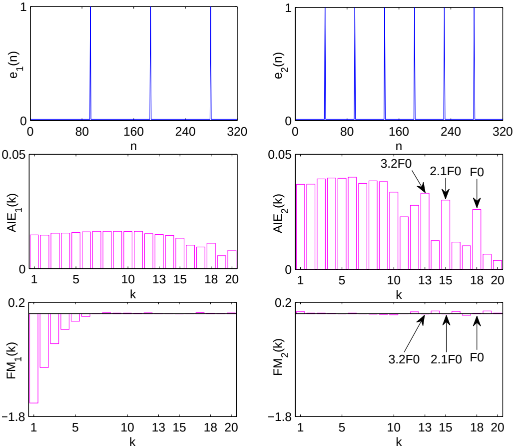
$f_c(18) = F0 = 172.4Hz$	$f_c(15) = 2.1F0$	$f_c(13) = 3.2F0$

Observation 1: pitch variation

AIE and FM for artificial excitation signals with different F0: $e_1(n)$ and $e_2(n)$.

F0: $e_1(n) = 86.2Hz$;

F0: $e_2(n) = 172.4Hz = f_c(18)$.



Observations:

- Amplitude peaks at F0 harmonics.
- Frequency in different bands.

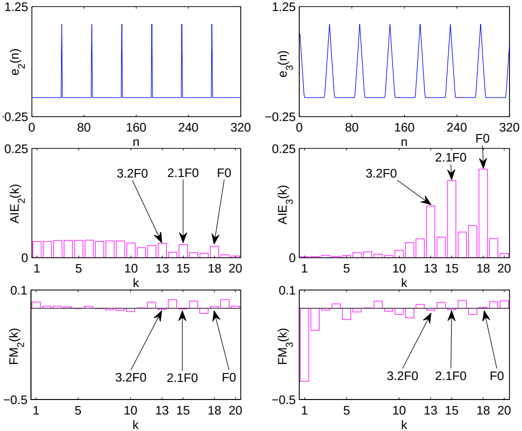
Observation 2: pitch epoch shape

AIE and FM for artificial excitation signals with different epoch shapes: $e_2(n)$ and $e_3(n)$.

Same F0.

Observations:

- Amplitude at F0 harmonics.
- Energy distribution across bands.



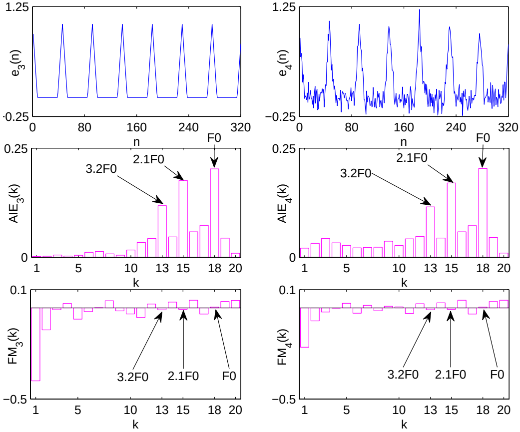
Observation 3: details between epochs

AIE and FM for artificial excitation signals with and without details between adjacent epochs: $e_3(n)$ and $e_4(n)$.

Same F0, same epoch shape.

Observations:

- Amplitude peaks at F0 harmonics.
- Energy in different bands.



Outline

Formant-related modulation properties

- Formant AM-FM modeling
  - Frequency and bandwidth tracking (Potamianos, 1996).
  - FM features (Ambikairajah, 2007).
- A bandlimited signal that describes a formant

$$F_k(n) = A_k(n) exp\left\{ j \left[\Theta_k(n) \right] \right\},$$

is characterized by two sequences:

  - $A_k(n)$ – Amplitude of formant;
  - $\Theta_k(n)$ – Phase of formant.

Primary speech components

A speech signal may contain:

- Formants and pitch harmonics: formulated as mono-component AM-FM terms $A(n)cos[\Theta(n)]$.
- Other components, *e.g.*, transitions between formants, interferences among harmonics and interactions within the vocal tract system, etc.

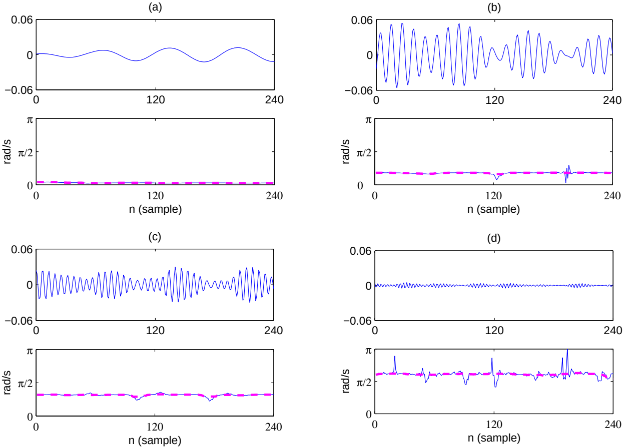
Thus, it can be written as a linear combination of AM-FM components which we called the primary speech components,

$$
\begin{aligned}
s(n) &= \sum_{k=1}^{K} A_k(n)cos\left[\Theta_k(n)\right] + \eta(n) \\
&= \sum_{k=1}^{K} A_k(n)cos\left\{\left[\Omega_c(k)n + \sum_{r=1}^{n} q_k(r)\right]\right\} + \eta(n).
\end{aligned}
$$

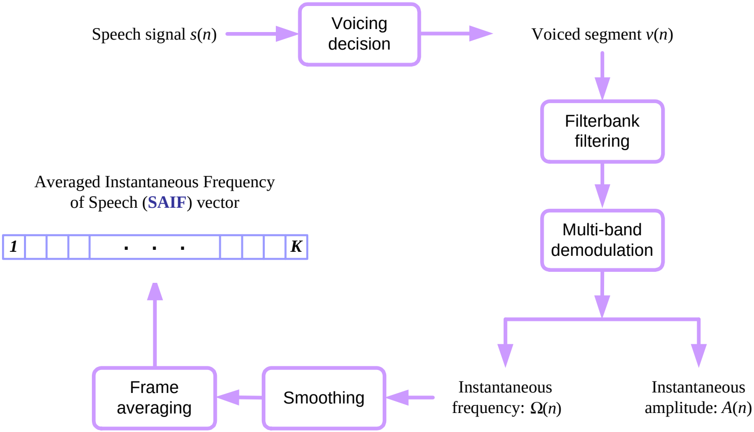
Frequency components in speech

Subband signals and their instantaneous frequency (IF) sequences.

- Centers:
  - (a). 0.03π;
  - (b). 0.18π;
  - (c). 0.31π;
  - (d). 0.61π.



Instantaneous frequency-based features



Feature analysis

- Artificially generated speech sounds: $s_1(n)$ and $s_2(n)$.
  - F0 and formant frequencies are specially settled.

$f_c(36) = 172.4Hz = F0$	$f_c(30) = 2.1F0$	$f_c(26) = 3.2F0$
$f_c(22) = 773.8Hz = F1$	$f_c(17) = 1161.8Hz = F2$	$f_c(7) = 2446.2Hz = F3$

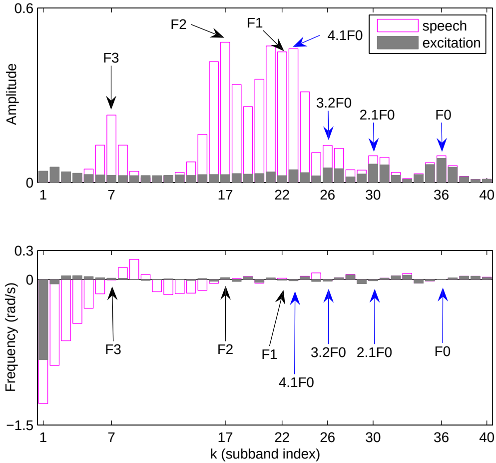
  - Formant bandwidth varies.
    - $s_1(n)$: $10Hz$
    - $s_2(n)$: $200Hz$

Observation 1: formants, harmonics and their interactions

Amplitude and frequency of primary speech components: $s_1(n)$.

Observations:

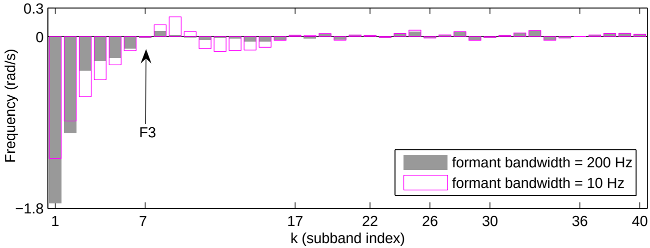
- Amplitude and frequency at formants and F0 harmonics.
- Harmonics noticeable in the lower frequency bands.
- Formants dominant in the higher frequency bands.



Observation 2: Formant bandwidth effect

Frequency of primary speech components: $s_1(n)$ and $s_2(n)$.

- FM around formants:
  - The one assuming larger bandwidth, less peaky formants, produces smaller FM components around the formants.



Outline

Feature evaluation: experimental set-up

- Feature:
  - MFCCs
  - Modulation features: RAIE, RAIF or SAIF
- Database: CU2C (Dual conditional speech corpus, 50 male speakers), NOISEX-92 (Noise database)
- Training data: Microphone speech (CU2C)
- Test data:
  - Matched condition: Microphone speech (CU2C)
  - Mismatched conditions:
    1. Microphone speech (CU2C) + additive noise (NOISEX-92)
    2. Telephone speech (CU2C)
- Modeling: 256 mixtures-GMM
- Score fusion of individual features:

$$\begin{aligned} & score &= \; & w_M \times score_M + w_R \times score_R \quad (w_M + w_R = 1) \\ \text{or} \quad & score &= \; & w_M \times score_M + w_S \times score_S \quad (w_M + w_S = 1) \end{aligned}$$

Evaluation 1: excitation modulation features

Feature configuration		IDER (%)	EER (%)
Baseline	MFCC	2.44	1.52
Effects of feature dimension	RAIE_*20*	40.72	13.17
	RAIF_*20*	35.11	10.42
	RAIE_*40*	27.28	9.46
	RAIF_*40*	19.67	8.01
	RAIEF_*40*	22.50	8.17
Combination with MFCC	MFCC+RAIE_*20*	2.39	1.49
	MFCC+RAIF_*20*	2.28	1.24
	MFCC+RAIE_*40*	2.44	1.44
	MFCC+RAIF_*40*	**2.06**	**1.27**
	MFCC+RAIEF_*40*	2.17	1.36

Evaluation 2: phase-related parameters

Feature configuration		IDER (%)	EER (%)
Baseline	MFCC	2.44	1.52
Effects of feature dimension	SAIF_*20*	6.33	3.72
	SAIF_*40*	**4.78**	**2.70**
Effects of frame length	SAIF_*bil_40*	6.39	3.64
	SAIF_*tril_40*	9.22	4.54
Supplementing cepstral features	MFCC+SAIF_*40*	1.83	1.16
	MFCC+SAIF_*bil_40*	1.89	1.21
	MFCC+SAIF_*tril_40*	1.78	1.33
Combining with source features	RAIE_*40*+SAIF_*40*	5.33	2.69
	RAIF_*40*+SAIF_*40*	4.61	2.65

Analysis 1: on individual features

- MFCCs *vs.* modulation features
  - MFCC features perform much better than the modulation features.
  - The best performed modulation feature has comparable dimension with MFCCs.
- Comparisons among modulation features
  - Dimension: High (40) > Low (20).
  - Configuration: AIF > AIE.
  - Information source: SAIF > RAIF.

Analysis 2: on fusion results

- Recognition accuracy
  - Modulation parameter sets when fusing with MFCCs can produce improved results.
  - Better performed modulation feature + MFCCs produce better fusion results.
- Complementary effects in fusion:

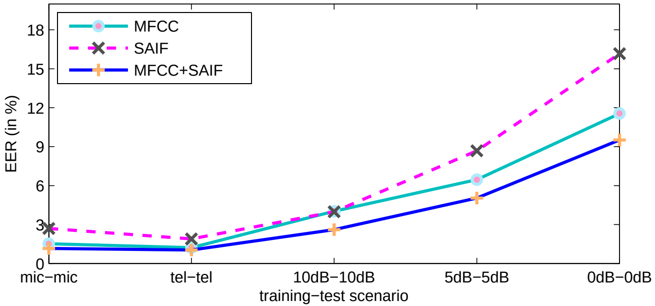
  SAIF + MFCCs > RAIE/RAIF + MFCCs > SAIF + RAIE/RAIF.

Phase information in speech signals can help MFCC-based SR system to achieve higher performance.

Outline

Speaker discrimination under adverse conditions

- Under clean and matched training-test conditions:



- Under mismatched conditions:

Training data	Test data	MFCC	SAIF_*40*	MFCC + SAIF_*40*
Mic.: clean	Mic.: 10dB	17.41	23.47	15.93
Mic.: clean	Tel.	18.94	24.67	18.35

Amplitude-modulated speech components

Dominant frequency of a subband signal relatively stable for a short time interval

Bandwidth, spectral envelop of a subband signal depends on the amplitude

AM – FM component of speech

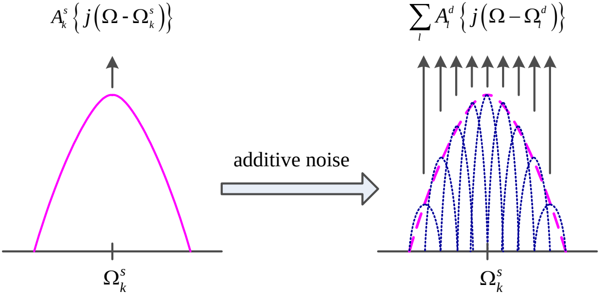
$$A_k^s(n)\cos\left[\Omega_c(k)n+\sum_{r=1}^{n} q_k(r)\right]$$

Take the estimate of dominant frequency, i.e., Ω_k^s, as instantaneous frequency in the concerned subband.

AM component

$$A_k^s(n)\cos\left(\Omega_k^s n\right)$$

Effects of additive noise



- A noise signal contains a large number of narrow-band frequency components.
- Around a specific speech component, *e.g.*, the kth, there are several affecting noise bands exist.

The slowly varying noise components around Ω_k^s generally distort its distribution by moving its mean towards $\Omega_c(k)$ and reducing its variance.

Effects of convolutive noise

In telephone networks, a speech signal $s(n)$ is transmitted through a channel with impulse response $c(n)$.

For the kth subband:

Assumption: $c_k(n) \quad = \quad A_k^c \delta(n - n_k^c)$

Output signal:

$$
\begin{aligned}
z_k(n) &= s_k(n) \otimes c_k(n) \\
&= \left\{ A_k^s(n) cos(\Omega_k^s n) \right\} \otimes \left\{ A_k^c \delta(n - n_k^c) \right\} \\
&= \underbrace{\left\{ A_k^c A_k^s(n - n_k^c) \right\}}_{\text{Amplitude}} \cdot \underbrace{\left\{ cos(\Omega_k^s n - \Omega_k^s n_k^c) \right\}}_{\text{Phase}}
\end{aligned}
$$

Brief introduction to feature mapping

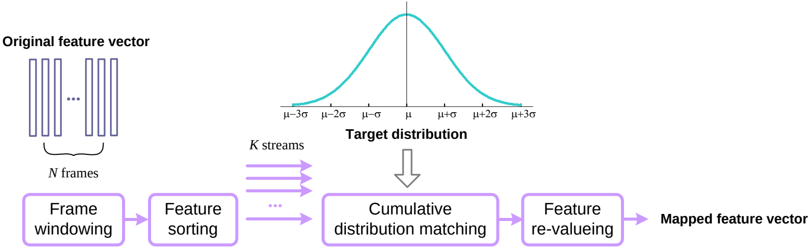
Feature mapping for speaker verification

- Proposed by Jason Pelecanos and Sridha Sridharan ("Feature Warping for Robust Speaker Verification", *2001: A Speaker Odyssey*).
- A post-processing method for robust feature extraction.
- Requires no noise statistics or channel models.

It is found that feature mapping method works well for cepstral features.

Then, what is a proper mapping process for phase-related parameters?

Mechanism of feature mapping



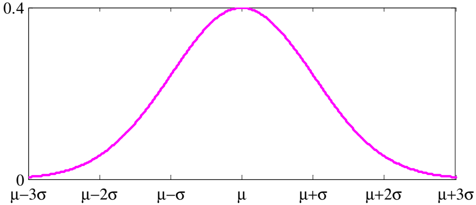
- **Frame windowing**: a sliding window to isolate N frames of speech features.
- **Feature sorting**: feature values in each stream sorted descendingly assuming rank 1 to N.
- **Cumulative distribution matching**: find the *relative* position for the feature in the target distribution

$$\int_1^u f_m(x)dx = \int_1^v f_t(y)dy,$$

u, v: ranks of the current feature in the measured and target distributions.

- **Feature re-valuing**: find the *absolute* feature value of the mapped parameter.

Feature-specific target distributions



MFCCs:

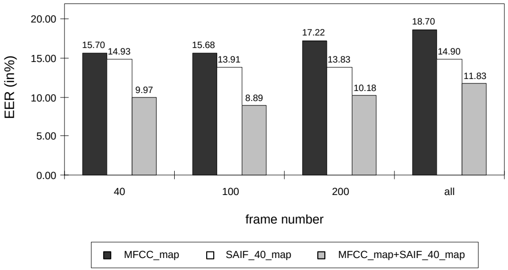
- Measured distribution: multi-modal in nature.
- All feature streams share the same target distribution: $N(0,1)$.

SAIF: *Histograms*

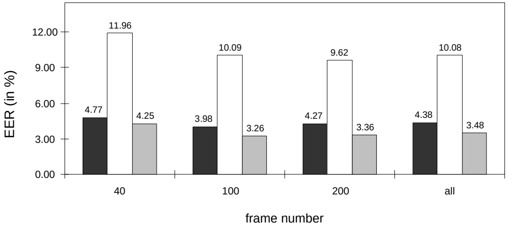
- Measured distribution: uni-modal.
- Individual target for each feature stream: $N(\mu, \sigma^2)$. For the kth parameter:
  - μ: $\Omega_c(k)$
  - σ: bandwidth $(ERB(k))/6$

Experimental results

(a). mismatched noise condition



(b). mismatched channel condition



Benchmark results:

	MFCC	SAIF_*40*	MFCC+SAIF_*40*
(a)	17.41	23.47	15.93
(b)	18.94	24.67	18.35

- Two mismatched conditions: noise or channel mismatch.
- Four kinds of window size for feature mapping:
  - 40 frame: 0.4s
  - 100 frame: 1s
  - 200 frame: 2s
  - all frame: length of the utterance

Observations and analysis

- Recognition performance
  - Performance of individual MFCC and SAIF features improved.
  - Combination of the magnitude-based and phase-related parameters demonstrate advantages.
- Feature mapping effectiveness
  - Under channel mismatch condition, feature mapping leads to larger improvement, especially for MFCCs.
  - For additive noise condition, SAIF features get more benefit from the mapping.

Outline

Summary

- Speaker-specific characteristics are identified in primary amplitude and frequency components of decomposed speech signal.
- Modulation feature vectors are generated by extracting the most dominant components in multi-band time-varying modulation parameters.
- Complementary assistance of phase and vocal excitation modulation features to MFCCs are confirmed by improvements of SR performance.
- Robustness of speaker parameters in various scenarios are enhanced by feature-dependent mapping approach.

Future directions

For robust speaker recognition,

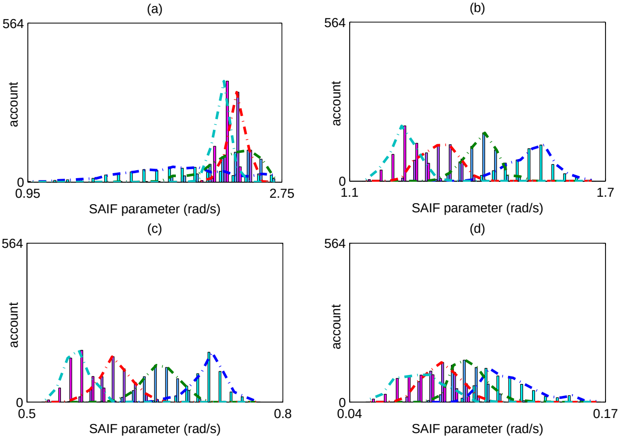
- Producing flexible forms for higher efficient features.
  e.g., weighting scheme among streams in a vector, selection criterion.
- Generating features from unvoiced segments.

For others,

- Modeling source signal of speech.
- Learning to recognize/predict human emotional states.
  - Application scenarios: human-computer interaction (HCI), interactive learning system, intelligent robotics, emotional state monitoring, etc.
  - Fusion with other modalities: video, EEG & physiological signals, etc.

Thank you very much!

Histogram statistics of SAIF streams

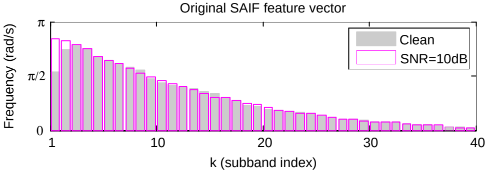


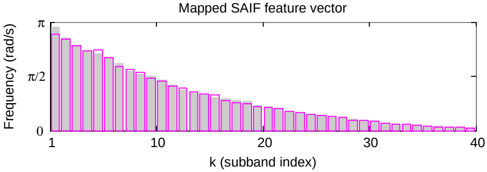
(a). $k = 1, \ldots, 4$, (b). $k = 10, \ldots, 13$, (c). $k = 20, \ldots, 23$, and (d). $k = 37, \ldots, 40$.

Target distribution

Original *vs.* mapped parameters

SAIF feature vectors:





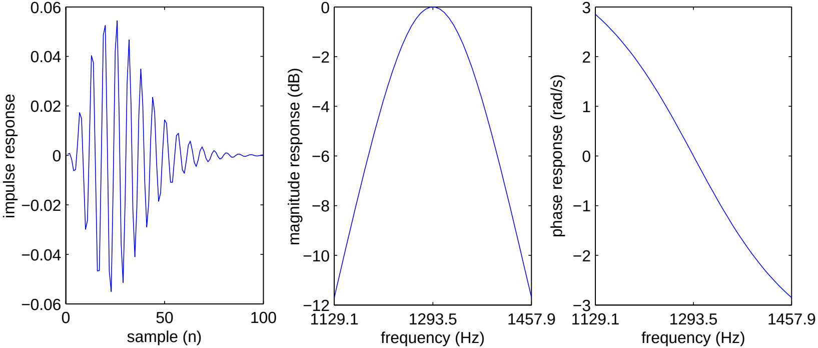
Observations:

- Before mapping:
  Difference is obvious in the high frequency region.

- After mapping:
  Difference in the high frequency region is mitigated, while they are small among other regions.

Gammatone filter-bank

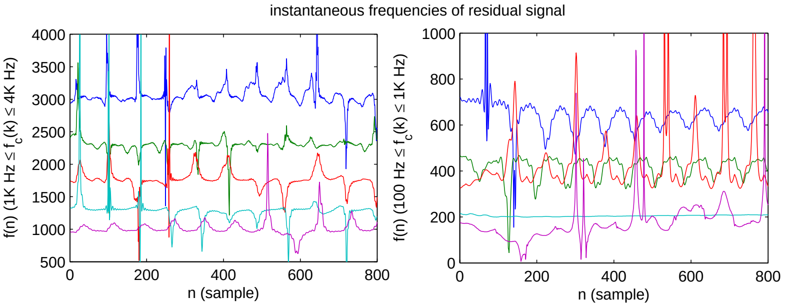
- Model the cochlea by a bank of overlapping bandpass filter.
- Filters are evenly-distributed on an ERB scale.
- The 4th filter among the 10: $f_c(4) = 1293.5Hz$, $bw(4) = 164.4Hz$.



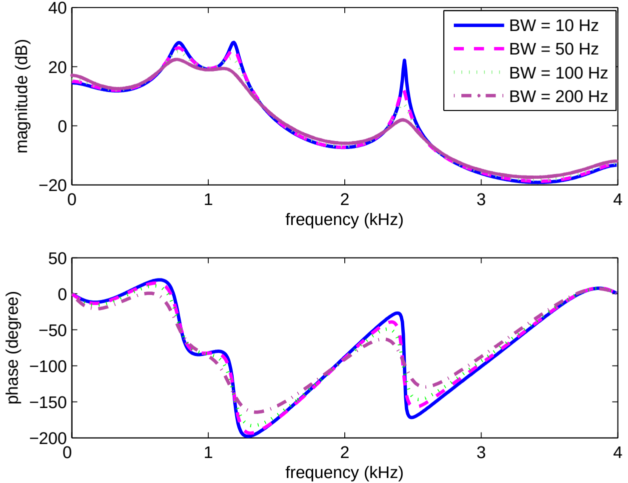
IF in excitation signal

Table: Center frequencies and ERB bandwidths of a 10-channeled Gamma-tone filter bank spaced on [100Hz, 4k Hz] (in Hz).

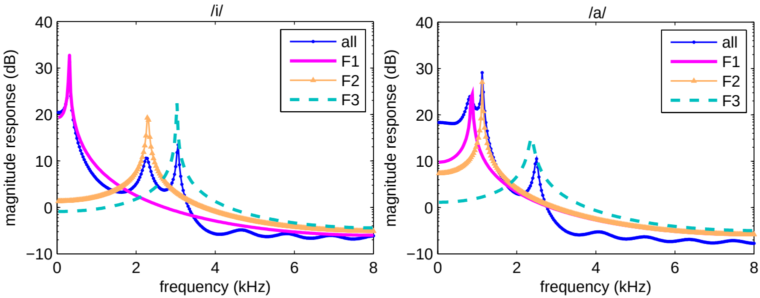
k	1	2	3	4	5	6	7	8	9	10
$f_c(k)$	3046.8	2308.5	1736.5	1293.5	950.4	684.6	478.7	319.2	195.7	100
$ERB(k)$	353.8	274.0	212.2	164.4	127.3	98.6	76.4	59.2	45.8	35.5



Formant bandwidth effect: LP spectra



Vocal tract resonance



The complex conjugate poles correspond to a vocal tract resonance is

$$z_k, z_k^* = exp\big(-\frac{\sigma_k}{f_s}\big)exp\big(\pm j\frac{2\pi}{f_s}F_k\big).$$

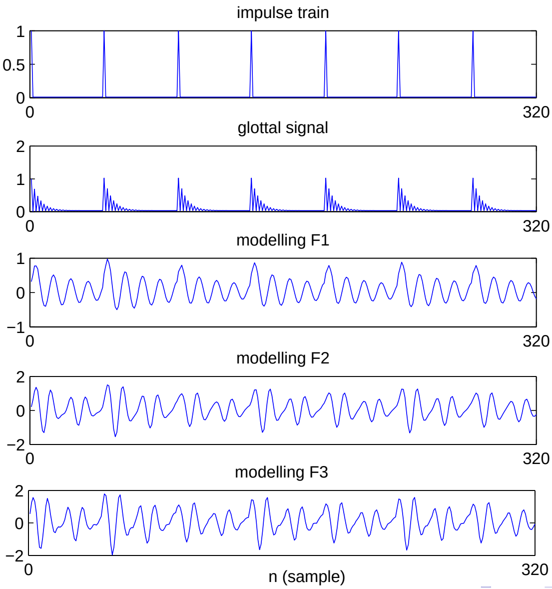
In the z transform domain, the bandwidth is determined by the radius of the poles, i.e.,

$$r_k = |z_k| = exp\big(-\frac{\sigma_k}{f_s}\big),$$

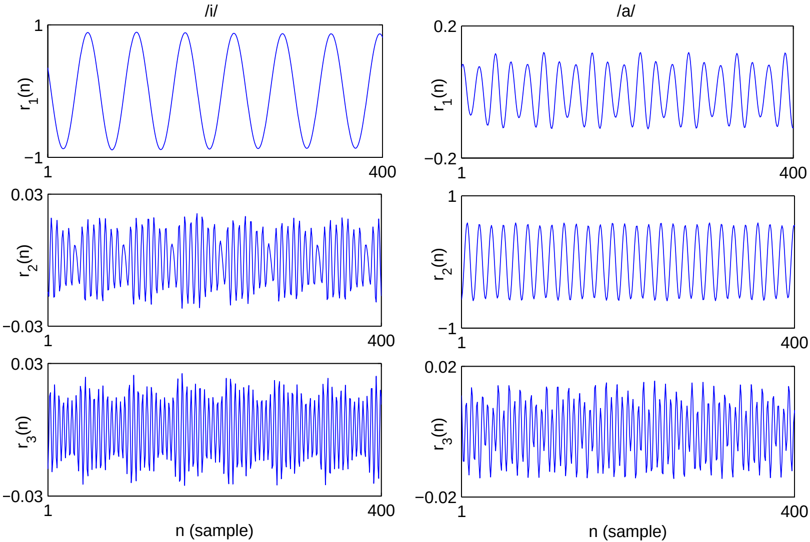
while, the angles of the conjugate poles from the origin are

$$\theta_k = \pm\frac{2\pi}{f_s}F_k.$$

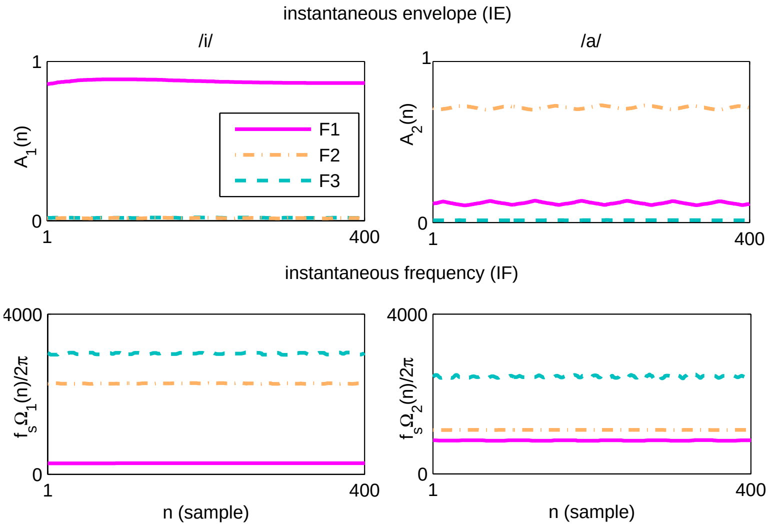
Voiced sound synthesis



Waveform of resonant signals



Instantaneous envelope and frequency of resonant signals



SID & SV tests

- SID tests
  - Close-set.
  - 1800 tests for each speaker.
- SV tests
  - 36 claimant tests.
  - 1764 imposter tests for each speaker.