Speaker Recognition

Digital Speech Processing in Noisy Environments

048831

Source: Thomas F.Quatieri "Discrete Time Speech Signal Processing"- chapter 14

Introduction

Speaker Identification

- Goal: to decide which voice model from a known set of voice models best characterizes a speaker

Speaker Verification

- Goal: to decide whether a speaker corresponds to a particular known voice (claimant) or to some other unknown voice (imposter)

claimant = target speaker
imposter = background speaker

- 2 types of error:
  - false acceptance – an imposter is accepted as a claimant
  - false rejection – claimant is rejected as imposter

Introduction

1. The first step: to build a model of the voice (target speaker + collection of background speakers).

2. Second stage: to match the features measured from the waveform of a test utterance against speaker models obtained during training.

3. Decision: target speaker or background speaker.

Spectral Features for Speaker Recognition

Mel-Cepstrum

1. STFT: $$X(n,\omega_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega_k m}$$

$\omega_k = \frac{2\pi}{N}k$, **N the DFT length**

2. Mel - scale filter bank $V_l(\omega)$

Example of filter bank

Series of filter frequency responses whose center frequencies are linear for low frequency and logarithmically increase with increasing frequency.

3. Energy: $$E_{mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l}\left|V_l(\omega_k)X(n,\omega_k)\right|^2$$

where L_l **and** U_l **the lower and upper frequency indices over which each filter** $V_l(\omega)$ **is nonzero and where:**

$$A_l = \sum_{k=L_l}^{U_l}\left|V_l(\omega_k)\right|^2$$

Spectral Features for Speaker Recognition

Mel-Cepstrum

The real cepstrum associated with $E_{mel}(n,l)$ computed for the speech frame at time n:

$$C_{mel}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} \log\{E_{mel}(n,l)\}\cos(\frac{2\pi}{R} lm)$$

where R is the number of filters and where we have used the even property of the real cepstrum to write the inverse transform in terms of the cosine basis.

Sub-Cepstrum

Convolution of the mel-scale filter impulse responses directly with the waveform x[n]. The result of this convolution can be expressed as:

$$\tilde{X}(n,\omega_l) = x[n] * v_l[n]$$

where $v_l[n]$ denotes the impulse response corresponding to the frequency response $V_l(\omega)$ of the lth mel-scale filter centered at frequency ω_l .

Spectral Features for Speaker Recognition

Sub-Cepstrum

The energy of the output of the l th subband filter can be taken as simply $\left|\tilde{X}(n,\omega_l)\right|^2$ or as a smoothed version of $\left|\tilde{X}(n,\omega_l)\right|^2$ over time:

$$E_{sub}(n,l) = \sum_{m=-N/2}^{N/2} p[n-m]\left|\tilde{X}(m,\omega_l)\right|^2$$

using a smoothing filter p[n].

Subband cepstrum is written as:

$$C_{sub}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} \log\{E_{sub}(n,l)\}\cos\left(\frac{2\pi}{R} lm\right)$$

Speaker Recognition Algorithms

Gaussian Mixture Model (GMM)

The subband filter energies more clearly show speech transitions, periodicity and short-time events, particulary in the high-frequency region.

In this example, the subband filter energies have not been smoothed in time.

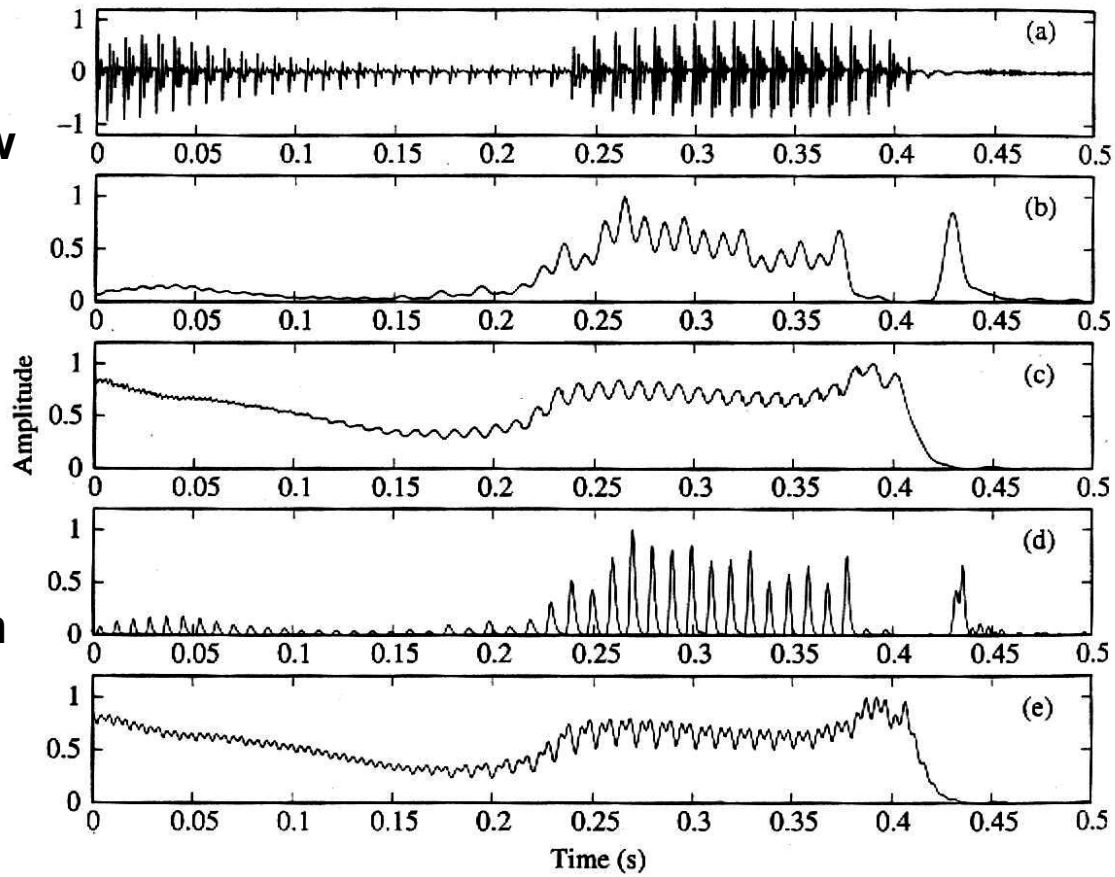


Figure 14.4 Energies from mel-scale and subband filter banks: (a) speech waveform; (b)–(c) mel-scale filter energy from filter number 22 (≈ 3200 Hz) and filter number 2 (≈ 200 Hz); (d)–(e) subband filter energy from filter number 22 (≈ 3200 Hz) and filter number 2 (≈ 200 Hz).

Speaker Recognition Algorithms

Minimum-Distance Classifier

One of the simplest approaches to speaker recognition is to compute the average of feature vectors over multiple analysis frames for speakers from testing and training data and then find the distance between these average test and training vectors. In speaker verification, we set a distance threshold below which we "detect" the claimant speaker; in identification, we pick the target speaker with smallest distance from that of the test speaker.

Speaker Recognition Algorithms

Minimum-Distance Classifier

Example: the average of mel-cepstral features for the test data:

$$\overline{C}_{mel}^{ts}[n] = \frac{1}{M}\sum_{m=1}^{M} C_{mel}^{ts}[mL, n]$$

the average of mel-cepstral features for the training data:

$$\overline{C}_{mel}^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C_{mel}^{tr}[mL, n]$$

where M is number of analysis frames and L is the frame length.

Speaker Recognition Algorithms

Minimum-Distance Classifier

The mean-squared difference:

$$D = \frac{1}{R-1}\sum_{n=1}^{R-1}(\overline{C}_{mel}^{ts}[n] - \overline{C}_{mel}^{tr}[n])^2$$

where R is the number of mel-cepstral coefficients.

In speaker verification:

if D < T, then speaker present.

In speaker identification:

The speaker that has minimum distance to the average feature vector of the test speaker.

Speaker Recognition Algorithms

Vector Quantization

A problem with the minimum-distance classifier is that it does not distinguish between acoustic speech classes. Individual speech events are blurred. We could do better if we average feature vectors over distinct sound classes(quasi-periodic, noise-like, impulse-like sounds, or even finer phonetic categorization.

That would reduce, for example, the phonetic difference in the feature vectors and help focus on speaker differences.

Speaker Recognition Algorithms

Vector Quantization

Example: the average of mel-cepstral features for the test data for the i th class:

$$\overline{C}_i^{ts}[n] = \frac{1}{M}\sum_{m=1}^{M} C_i^{ts}[mL, n]$$

the average of mel-cepstral features for the training data:

$$\overline{C}_i^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n]$$

where M is number of analysis frames and L is the frame length.

Speaker Recognition Algorithms

Vector Quantization

The mean-squared difference for each class:

$$D(i) = \frac{1}{R-1} \sum_{n=1}^{R-1} (\overline{C}_i^{ts}[n] - \overline{C}_i^{tr}[n])^2$$

Finally, we average over all classes as:

$$D(I) = \frac{1}{I} \sum_{i=1}^{I} D_i$$

where *I* is the number of classes.

Speaker Recognition Algorithms

Gaussian Mixture Model (GMM)

The probability of a feature vector being in any one of I acoustic classes for a particular model, denoted by λ is represented by the mixture of different Gaussian pdfs:

$$p(x \mid \lambda) = \sum_{i=1}^{I} p_i b_i(\underline{x})$$

where the $b_i(\underline{x})$ are the component mixture densities and p_i are the mixture weights.

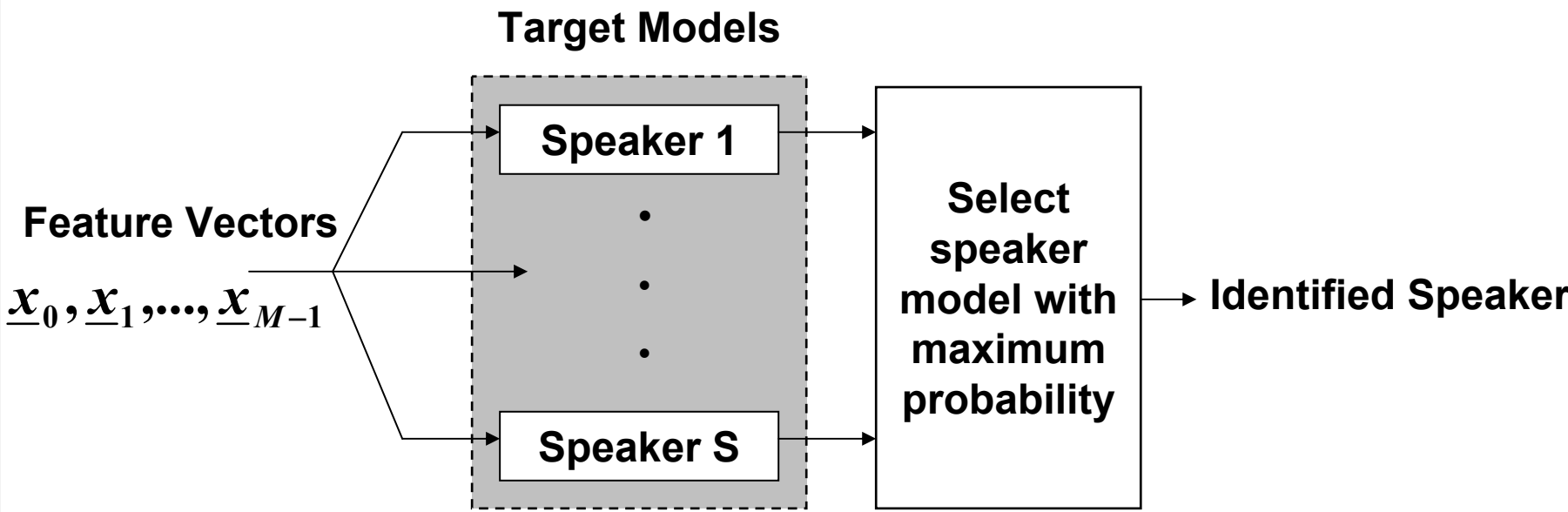
The speaker model λ then represents the set of GMM mean, covariance and weight parameters:

$$\lambda = \{p_i, \underline{\mu}_i, \sum_i \}$$

Speaker Recognition Algorithms

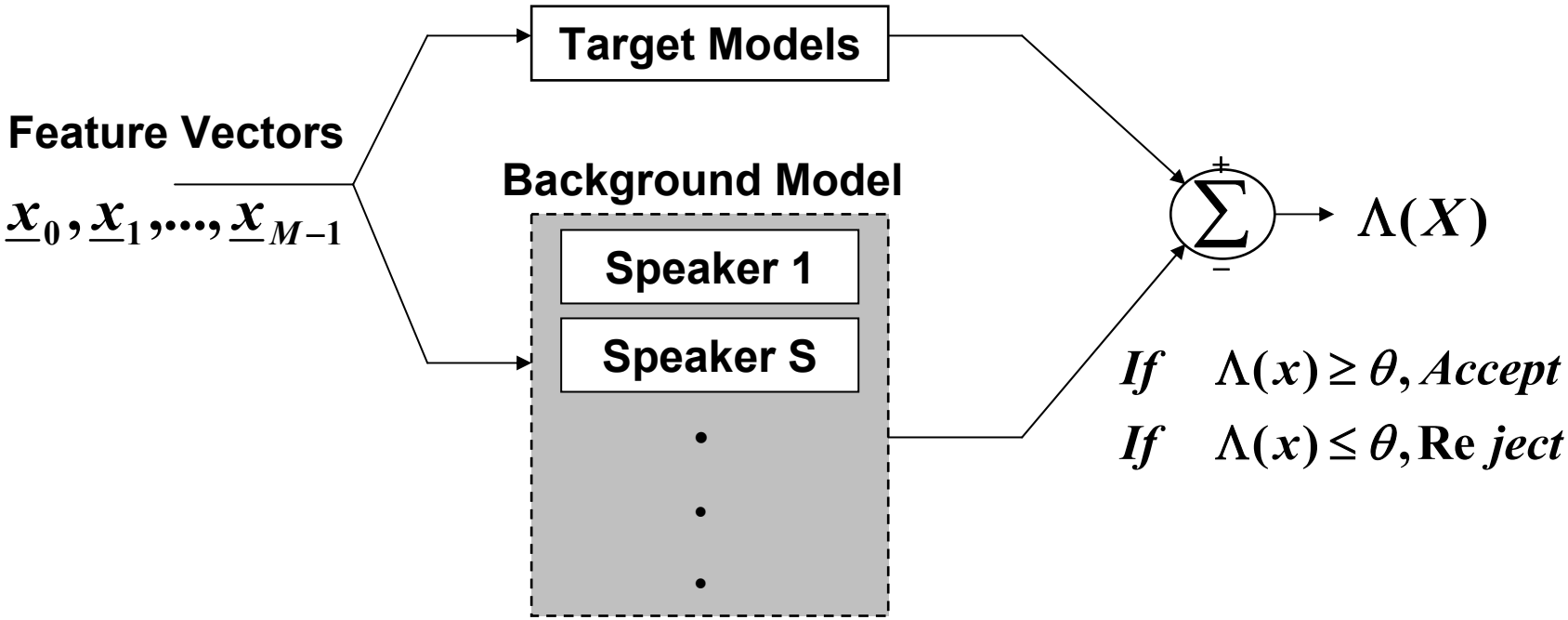
Gaussian Mixture Model (GMM)

Speaker Identification



Speaker Recognition Algorithms
Gaussian Mixture Model (GMM)

Speaker Verification



$$\Lambda(X) = \log[p(X \mid \lambda_C)] - \log[p(X \mid \lambda_{\overline{C}})]$$

λ_c - The claimed speaker model

Speaker Recognition Algorithms

Gaussian Mixture Model (GMM)

TIMIT - speech database, recorded in quiet environment, 630 speakers

NTIMIT - TIMIT transmitted over actual telephone channels

Results of GMM speaker identification system

- The speaker identification performance for clean speech (TIMIT) is near 100% up to a population size of 630 speakers
- The speaker identification performance with telephone speech drops off considerably as population size increases

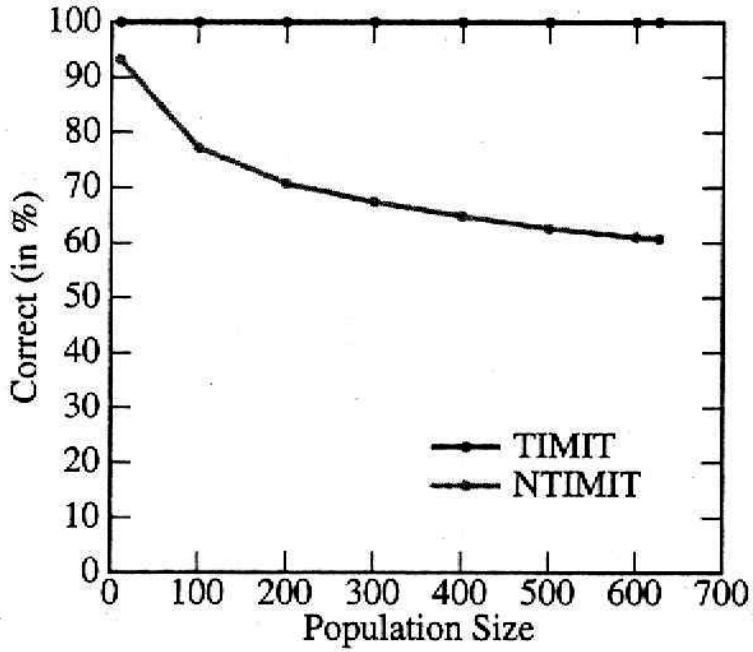


Figure 14.8 Performance of GMM speaker identification system on TIMIT and NTIMIT databases, as a function of number of speakers. Performance is measured as percent correct.

SOURCE: D.A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models" [68]. ©1995, *MIT Lincoln Laboratory Journal*. Used by permission.

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

The glottal flow derivative appears to be speaker-dependent. For example, the flow can be smooth, as well the folds never close completely, corresponding perhaps to a "soft" voice, or discontinuous, as when they close rapidly, giving perhaps a "hard" voice. The flow at the glottis may be turbulent, as when air passes near a small portion of the folds that remains partly open. When this turbulence, referred to as aspiration, occurs often during vocal cord vibration, it results in "breathy" voice.

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

The features we want to capture through the coarse structure include the glottal open, closed, and return phases, the speeds of opening and closing, and the relationship between the glottal pulse and the peak glottal flow. To model the coarse component of the glottal flow derivative we use the Liljencrants-Fant (LF) model:

$$v_{LF}(t) = \begin{cases} 0, & 0 \le t < T \\ E_0 e^{\alpha(t-T_0)} \sin[\Omega_0 (t - T_0)], & T_0 \le t < T_e \\ -E_1 [e^{-\beta(t-T_e)} - e^{-\beta(T_c - T_e)}], & T_e \le t < T_c \end{cases}$$

where:

$$E_1 = \frac{E_e}{1 - \exp[-\beta(T_c - T_e)]} \qquad E_0 = \frac{E_e}{e^{\alpha(T_e - T_0)} \sin[\Omega_0 (T_e - T_0)]}$$

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

Description of seven parameters of the LF model for glottal flow derivative waveform:

Table 5.1 Description of the seven parameters of the LF model for the glottal flow derivative waveform.

SOURCE: M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification" [19]. ©1999, IEEE. Used by permission.

T_o	The time of glottal opening.
α	Factor that determines the ratio of E_e to the peak height of the positive portion of the glottal flow derivative.
Ω_o	Frequency that determines flow derivative curvature to the left of the glottal pulse; also determines how much time elapses between the zero crossing and T_e.
T_e	The time of the maximum negative value of the glottal pulse.
E_e	The value of the flow derivative at time T_e.
β	An exponential time constant which determines how quickly the flow derivative returns to zero after time T_e.
T_c	The time of glottal closure.

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

LF model for the glottal flow derivative wave form:

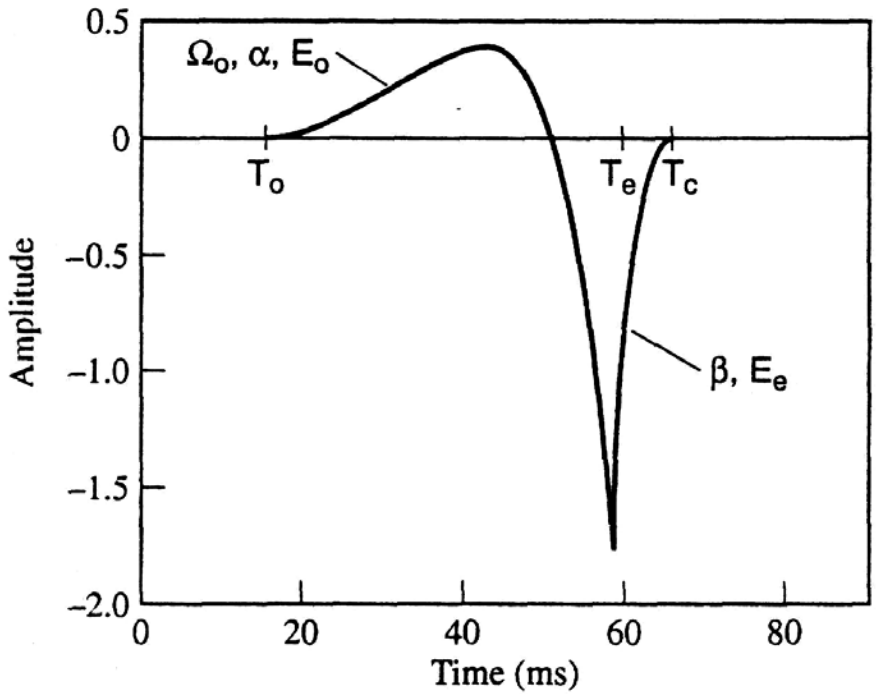


Figure 5.22 LF Model for the glottal flow derivative waveform.

SOURCE: M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification" [19]. ©1999, IEEE. Used by permission.

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

In characterizing the glottal flow derivative we define five time intervals within a glottal cycle. The first three intervals correspond to timing within the LF model of coarse structure, while the last two intervals come from timing mesurements made on formant modulation.

1. $I_1 = [0, T_o]$ is the closed phase for the LF model.
2. $I_2 = [T_o, T_e]$ is the open phase for the LF model.
3. $I_3 = [T_e, T_c]$ is the return phase for the LF model.
4. $I_4 = [0, T_f]$ is the closed phase for the formant modulation.
5. $I_5 = [T_f, T_e]$ is the open phase for the formant modulation.

Non -Spectral Features for Speaker Recognition

Glottal Flow Derivative

Example of a glottal flow derivative estimate and its coarse and fine structure:

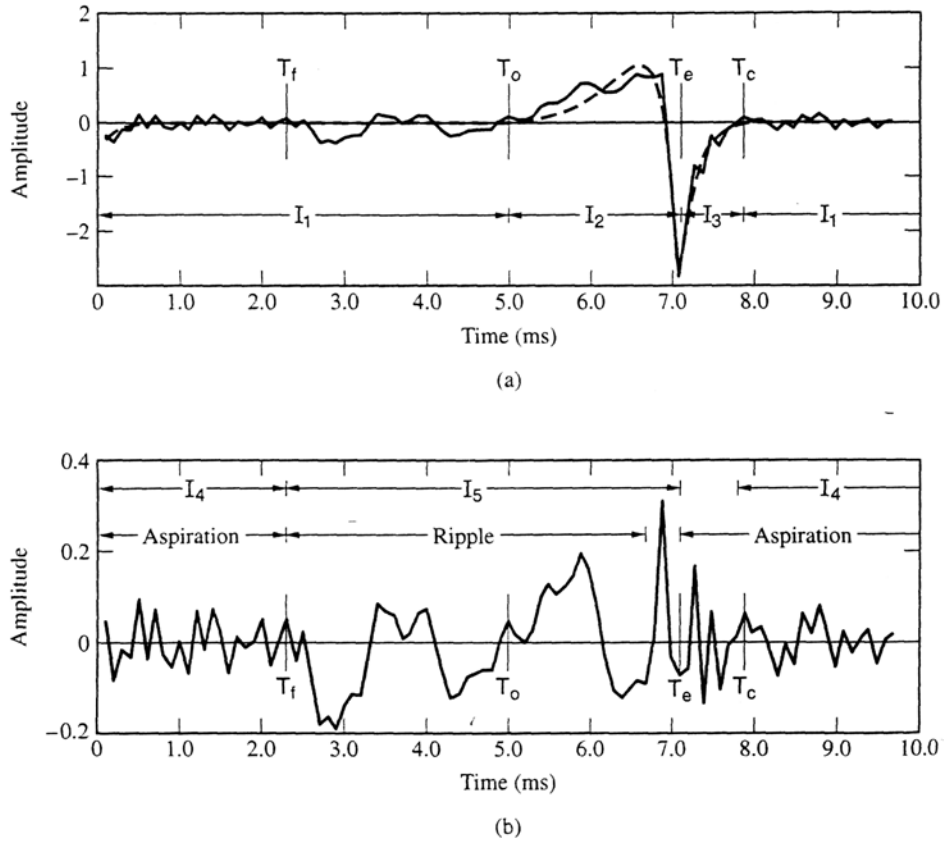


Figure 5.23 Example of a glottal flow derivative estimate and its coarse and fine structure: (a) estimated glottal flow derivative (solid) and overlaid LF model, i.e., the coarse structure (dashed); (b) the fine structure obtained by subtracting the coarse structure from the glottal flow derivative estimate. Aspiration and ripple are seen in different intervals of the glottal cycle.

Non -Spectral Features for Speaker Recognition

Source Onset Timing

The glottal flow features were chosen based on common properties of the glottal source. But actually there are numerous atypical cases. One example is multiple points of excitation within a glottal cycle. Such secondary pulses were observed to occasionally "excite" formants different from those corresponding to the primary excitation. The presence of such secondary source pulses may in part explain the improved speaker identification scores achieved by measuring pulse onset times in different formant bands and appending these onset times as features to mel-cepstral features.

Signal Enhancement for the Mismatched Condition

Training and test data may experience different distortions, a scenario we refer to as the mismatched condition, and this can lead to far greater performance loss than for matched condition in degrading environments.

Channels are characterized by:

- **additive noise**
- **linear distortions**
- **nonlinear distortions**

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

Signal enhancement methods for reducing linear channel distortions:

- **CMS - Cepstral Mean Substraction**
- **RASTA - RelAtive SpecTrAl processing**
- **DCC - Delta Cepstral Coefficients**

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

CMS

The sequence x[n] is passed through a linear time-invariant channel distortion g[n]:

$$y[n] = x[n] * g[n]$$

The logarithm of the STFT magnitude can be approximated as:

$$\log|Y(n,\omega)| \approx \log|X(n,\omega)| + \log|G(\omega)|$$

where w[n] is long and smooth relative to g[n].

As a function of time at each frequency ω, the channel frequency response G(ω) is seen as a constant disturbance. Assuming that the speech component log|X(n, ω)| has zero mean in the time dimension, we can estimate and remove the channel disturbance while keeping the speech contribution intact.

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

RASTA

A generalization of CMS is RelAtive SpecTrAl processing (RASTA) of temporal trajectories. RASTA addresses the problem of a slowly time-varying linear channel g[n,m] (i.e., convolutional distortion) in contrast to the time-invariant channel g[n] removed by CMS. The essence of RASTA is a cepstral lifter that removes low and high modulation frequencies and not simply the DC component, as does CMS.

The modified STFT magnitude used in RASTA enhancement is:

$$\left|\hat{X}(n,\omega)\right| = \exp\left\{\sum_{m=-\infty}^{\infty} p[n-m]\log\left|Y(m,\omega)\right|\right\}$$

where a single filter p[n] is used along each temporal trajectory and where Y(n,ω) denotes the STFT of a convolutionally distorted sequence x[n].

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

After being modified by CMS or the RASTA filter p[n] the mel-scale filter log-energies become:

$$\log[\tilde{E}_{mel}(n,l)] = \sum_{m=-\infty}^{\infty} p[n-m]\log[E_{mel}(m,l)]$$

where

$$E_{mel}(n,l) = \sum_{k=L_l}^{U_l} \left| V_l(\omega_k) X(n,\omega_k) G(\omega_k) \right|^2$$

Assume

$$G(\omega_k) \approx G_l, \quad for \quad k = [L_l \quad U_l]$$

where G_l represents a constant level for the lth mel-scale filter over its bandwidth.

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

$$E_{mel}(n,l) = \sum_{k=L_l}^{U_l} \left| V_l(\omega_k) X(n,\omega_k) G(\omega_k) \right|^2$$

$$\approx \left|G_l\right|^2 \sum_{k=L_l}^{U_l} \left| V_l(\omega_k) X(n,\omega_k) \right|^2$$

$$\log[E_{mel}(n,l)] \approx \log\left|G_l\right|^2 + \underbrace{\log[\sum_{k=L_l}^{U_l} \left| V_l(\omega_k) X(n,\omega_k) \right|^2]}_{\text{desired log-energy}}$$

CMS and RASTA can now view the channel contribution to the mel-scale filter log-energies as an additive disturbance to the mel-scale filter log-energies of the speech component.

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

DCC

An important objective in recognition systems is to find features that are channel invariant. Dynamic features that reflect change over time can have this property.

One formulation of the delta cepstrum invokes the difference between two mel-cepstral feature vectors over two consecutive analysis frames.

$$\Delta \log[E_{mel}(pL,l)] = \log[E_{mel}(pL,l)] - \log[E_{mel}((p-1)L,l)]$$

For a time-invariant channel, the channel contribution is removed by the difference operation. The delta cepstrum has been found to contain speaker identifiability, with or without channel distortion, but does not perform as well as the mel-cepstrum in speaker recognition. The mel-cepstrum and delta cepstrum are ofen combined as one feature sets in order to improve recognition performance.

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

Speaker verification

TSID (Tactical Speaker Identification Database) consists of 35 speakers reading sentences, digits and directions over a veriety of very degraded and low bandwidth wireless radio channels, including cellular and push-to-talk. These channels are characterized by additive noise and linear and nonlinear distortions. For each such "dirty" recording, a low noise and high bandwidth "clean" reference recording was made simultaneously at the location of the transmitter.

The performance is given in the terms of the equal error rate (EER). The EER is the point along the DET where the % of false acceptance and % of false rejection errors are equal, and thus provides a concise means of performance comparison.

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

Speaker verification

The mismatched case of training on clean and testing on dirty data gives has an EER of approximately 50% (performance of flipping a coin).

CMS alone gives greater performance than DCC alone, while the combination of CMS and DCC yields greater performance than either sole compensation method.

Table 14.3 Speaker verification results with various channel compensation techniques using the TSID database [45]. Results using CMS, DCC, or combined CMS and DCC show ERR performance improvements in the Clean/Dirty mismatched case.

SOURCE: M. Padilla, *Applications of Missing Theory to Speaker Recognition* [45]. ©2000, M. Padilla and Massachusetts Institute of Technology. Used by permission.

Equalization	ERR(approx)
Clean/Dirty + CMS + DCC	23%
Clean/Dirty + CMS	28%
Clean/Dirty + DCC	43%
Clean/Dirty	49%
Dirty/Dirty	9%
Clean/Clean	4%

Signal Enhancement for the Mismatched Condition

Nonlinear Channel Distortion

Telephone handset nonlinearity often introduces resonances that are not present in the speech spectrum – "phantom formants".

The comparison of all-pole spectra from a TIMIT waveform and its counterpart carbon-button microphone version from HTIMIT :

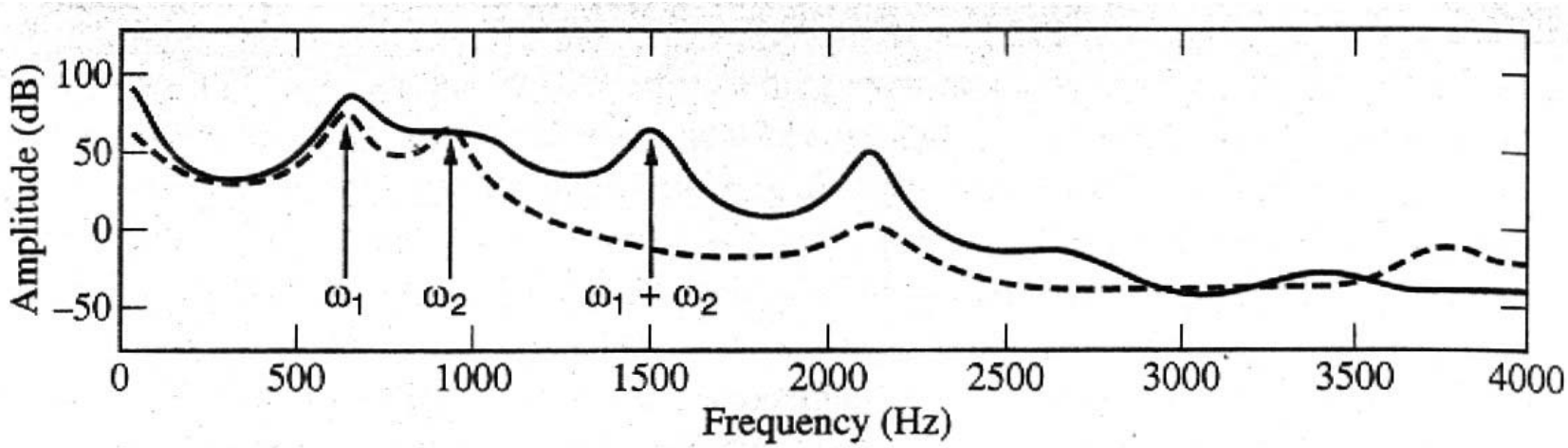


Figure 14.13 Illustration of phantom formants, comparing all-pole spectra from wideband TIMIT (dashed) and carbon-button HTIMIT (solid) recordings. The location of the first phantom formant ($\omega_1 + \omega_2$) is roughly equal to the sum of the locations of the first two original formants (ω_1 and ω_2). The (14th-order) all-pole spectra are derived using the autocorrelation method of linear prediction.

Signal Enhancement for the Mismatched Condition

Nonlinear Channel Distortion

One way for reducing handset mismatch between high- and low-quality handsets is handset mapper.

An example of mapping an electret handset to carbon-button handset output:

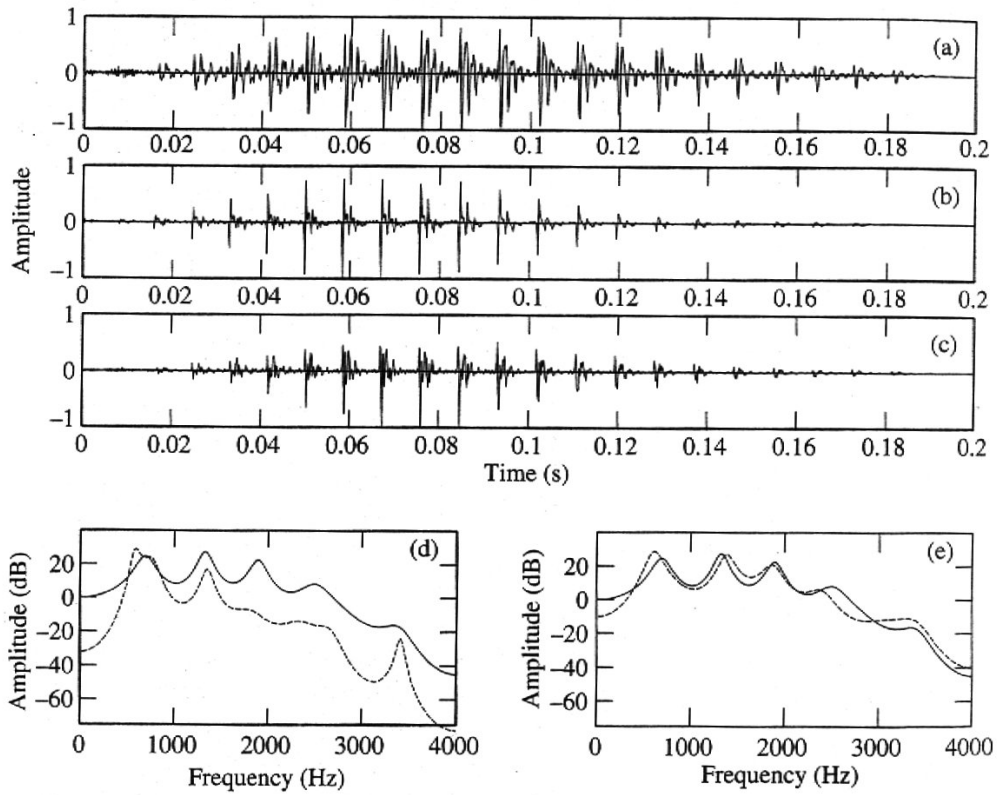


Figure 14.16 Example of electret-to-carbon-button mapping: (a) electret waveform output; (b) carbon-button waveform output; (c) electret-to-carbon mapped waveform; (d) comparison of all-pole spectra from (a) (dashed) and (b) (solid); (e) comparison of all-pole spectra from (b) (solid) and (c) (dashed). All-pole 14th order spectra are estimated by the autocorrelation method of linear prediction.

Signal Enhancement for the Mismatched Condition

Nonlinear Channel Distortion

The strategy is to assume two handset classes: a high-quality electret and a low-quality carbon-button handset.

We apply an electret-to-carbon-button nonlinear mapper to make electret speech utterances appear to come from carbon-button handsets when a handset mismatch occurs between the target training and test data.

		Test	
		ELEC	CARB
Train	CARB	No map	Map test to CARB
	CARB	Map test to CARB	No map

Signal Enhancement for the Mismatched Condition

Nonlinear Channel Distortion

Hnorm

The second approach to account for differences in handset type across training and test data modifies the likelihood scores, rather than operating on the waveform, as does the handset mapper approach.

Hnorm was motivated by the observation that, for each target model, likelihood ratio scores have different means and ranges for utterances from different handset types. The approach of Hnorm is to remove these mean and range differences from likelihood ratio scores. For each target model, we estimate the mean and standard deviation of the scores of imposter utterence from different handset types, and then normalize out the mean and variance to obtain a zero-mean and unity-variance score distribution.

For example, for carbon-button-handset type, during testing, we use the modified score:

$$\tilde{\Lambda}(\underline{x}) = \frac{\Lambda(\underline{x}) - u^{carb}}{\sigma^{carb}}$$

Signal Enhancement for the Mismatched Condition

Linear Channel Distortion

Speaker verification

In the matched condition, the target training and test data come from the same telephone number and the same handset type for target test trials, but different telephone numbers (and possibly different handset types) for imposter test trials.

In mismatched condition, training and test data come from different telephone numbers and thus possibly different handset types for both target and imposter test trials.

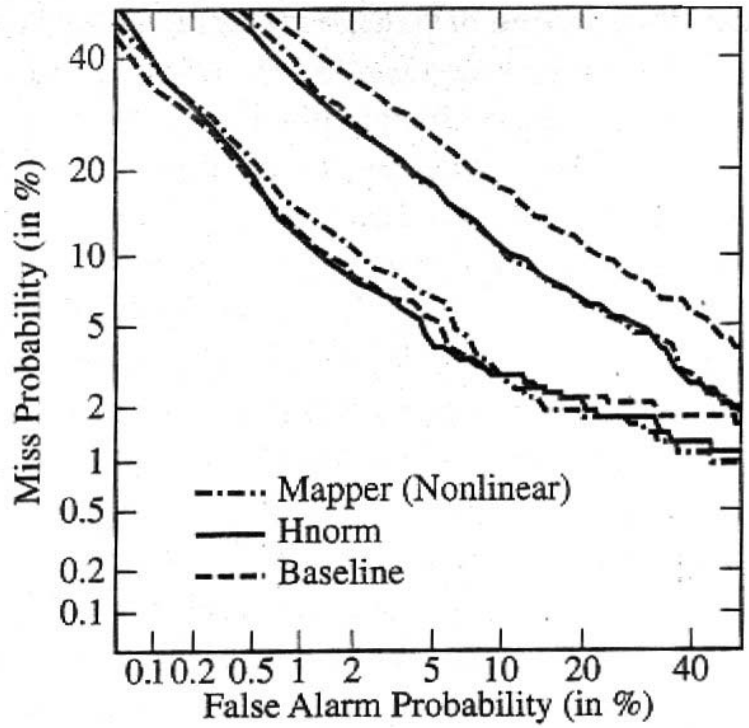


Figure 14.18 Comparison of DET performance using the nonlinear handset mapper and handset normalization of likelihood scores (hnorm) in speaker verification: baseline (dashed), hnorm (solid), and nonlinear mapper (dashed-dotted). Upper curves represent the mismatched condition, while lower curves represent the matched condition.

* **For either Hnorm or the mapper approach, we require a handset detector**

Speaker Recognition
Summary

- **Estimating features that characterize a speaker's voice:**
  - **Mel - sepstrum**
  - **Sub – sepstrum**
- **Approaches to speaker recognition:**
  - **Minimum - Distance Classifier**
  - **Vector Quantization**
  - **Gaussian Mixture Model (GMM)**
- **Linear channel compensation method:**
  - **CMS - Cepstral Mean Substraction**
  - **RASTA - RelAtive SpecTrAl processing**
  - **DCC - Delta Cepstral Coefficients**
- **Nonlinear channel compensation method:**
  - **Handset mapper**
  - **Hnorm**

Mel-scale filter bank

Series of filter frequency responses whose center frequencies are

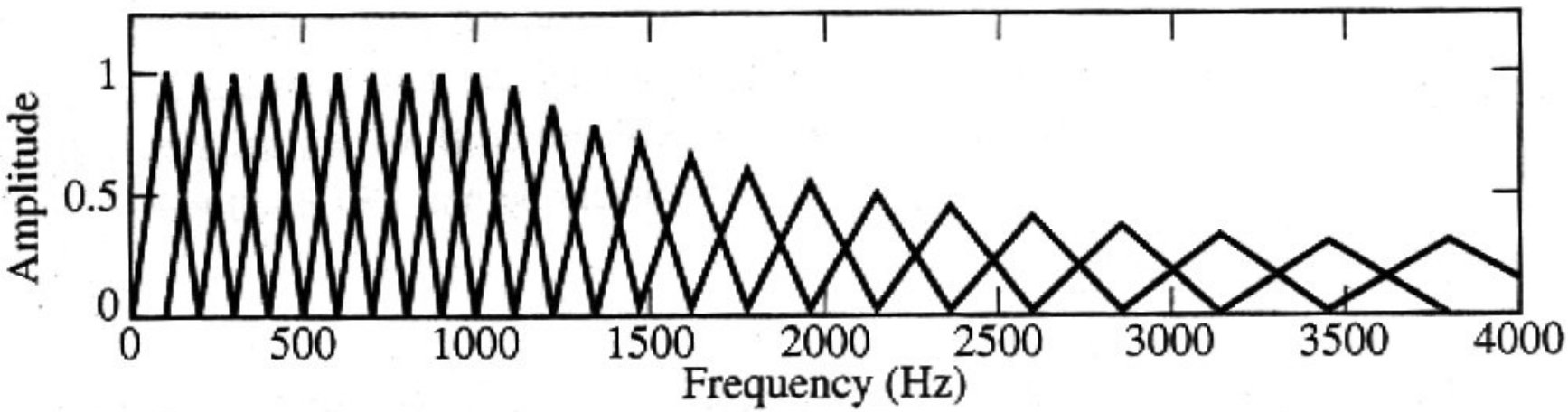


Figure 14.3 Triangular mel-scale filter bank used by Davies and Mermelstein [7] in determining spectral log-energy features for speech recognition. The 24 filters follow the *mel-scale* whereby band edges and center frequencies of these filters are linear for low frequency and logarithmically increase with increasing frequency, mimicking characteristics of the auditory critical bands. Filters are normalized according to their varying bandwidth.