CECS401
Fundamentals of Spoken Language Processing

Note-11
Tuesday 10/5/99
F. Speech Feature Analysis

Spectral distortion measure for automatic speech recognition

For automatic speech recognition, a distortion measure should be sensitive to spectral differences between sound classes and insensitive to differences within individual sound classes.

Relevant differences:

significant difference in formant locations.

Irrelevant differences:

gain, spectral tilt, pitch, distortion due to transmission channels and mismatch of transducers.
Mean-squared log spectral distance

\[ d(S(\omega), S'(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 d\omega \]

\[ = \sum_{m=-\infty}^{\infty} (c_m - c'_m)^2 \]

- gain is irrelevant
- \( c_m = c_{-m} \)
- the variance of \( c_m \) is decreasing with \( m \), so can use a sum of \( Q \) terms. For LPC cepstrum, \( p \leq Q \leq \frac{3}{2}p \).

Therefore the ms log spectral distance can be modified as

\[ \hat{d}(S(\omega), S'(\omega)) \triangleq 2 \sum_{m=-1}^{Q} (c_m - c'_m)^2 \]
Spectral-slope distance:

\[ d_{SS}(S(\omega), S'(\omega)) = \frac{1}{2\pi} \int_0^\pi \left| \frac{d \log S(\omega)}{d\omega} - \frac{d \log S'(\omega)}{d\omega} \right|^2 d\omega \]

\[ = \sum_{m=-\infty}^\infty (mc_m - mc'_m)^2 \]

For practical computation:

\[ \hat{d}_{SS}(S(\omega), S'(\omega)) \triangleq 2 \sum_{m=1}^Q m^2 (c_m - c'_m)^2 \]

- Since spectral slopes around spectral peaks (formants) are in general larger than at the spectral valleys, \( \hat{d}_{SS}(S, S') \) is sensitive to the differences of formant locations between \( S \) and \( S' \).
- The index weighting compensates for the \( \frac{1}{m^2} \) decay of cepstral variance with index \( m \).
**Bandpass-liftered cepstral distance:**

- Low-order $c_m$’s are sensitive to spectral tilts representing speaker glottal shape, transmission channel, etc.
- High-order $c_m$’s are sensitive to noise.
- The undesired sensitivities can be reduced by weighting the $c_m$’s with a tapering window

$$w(m) = \begin{cases} 1 + \frac{Q}{2} \sin \left( \frac{\pi m}{Q} \right) & 1 \leq m \leq Q \\ 0 & \text{otherwise} \end{cases}$$
The windowing in quefrency domain leads to the bandpass liftered cepstral distance:

\[
\hat{d}_{SS}(S(\omega), S'(\omega)) \overset{\triangle}{=} 2 \sum_{m=1}^{Q} (w(m)c_m - w(m)c'_m)^2
\]

\[
= 2 \sum_{m=1}^{Q} w^2(m)(c_m - c'_m)^2
\]

Example.

The effect of cepstral liftering on a log LPC spectrum.

- The liftered cepstral sequence \(w(m)c_m\) corresponds to a smoothed log spectrum.
- The LPC spectral tilt is removed.
Figure 4.20 Effects of cepstral liftering on a log LPC spectrum, as a function of the lifter length ($L = 8$ to 16) (after Juang et al. [11]).
Speech distortion measure for speech quality evaluation:
The need arises from quantitative evaluation of speech coders, speech synthesizers, and speech enhancement algorithms, to compare the machine generated speech $\hat{s}(n)$ with the original speech $s(n)$.
The distortion measure should be perceptually relevant and correlate with human subject evaluation tests.

Signal-to-noise ratio:

$$SNR = 10 \log_{10} \frac{E_s}{E_\epsilon} = 10 \log_{10} \frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}(n)]^2}$$

$SNR$ is not well related to any subjective attribute of speech quality. A deceptive high $SNR$ measure can be obtained if an utterance contains a high concentration of voiced segments, whereas noise has a stronger effect in low-energy segments (e.g., unvoiced fricatives, stops).
Segmental signal-to-noise ratio $SNR_{seg}$:
Segmental SNR is measured over short frames and the results averaged:

$$SNR_{seg} = \frac{1}{M} \sum_{j-1}^{M} 10 \log_{10} \left[ \frac{E_s(j)}{E_e(j)} \right]$$

where $j$ index frames.

$SNR_{seg}$ permits the objective measure to assign equal weight to loud and soft portions of speech.

Frequency-weighted segmental SNR ($SNR_{fw-seg}$):
Segmental SNR is measured in each critical band and is weighted according human perception sensitivity in each band:

$$SNR_{fw-seg} = \frac{1}{M} \sum_{j-1}^{M} 10 \log_{10} \left[ \frac{\sum_{k=1}^{K} w_{j,k} [E_{s,k}(j)/E_{e,k}(j)]}{\sum_{k=1}^{K} w_{j,k}} \right]$$
where \( k \) index critical bands, \( w_{j,k} \) denotes the perceptual weighting function.

\( SNR_{f_w-seg} \) is a better predictor of speech quality than \( SNR \) and \( SNR_{seg} \).

**Other measure:**

A family of Euclidean distances defined on different sets of LPC parameters.
Comparison of correlations between objective measure and subjective speech quality:

| Objective Quality Measure                          | $|\rho| $ |
|---------------------------------------------------|-------|
| SNR                                               | 0.24* |
| \(\text{SNR}_{\text{op}}\)                       | 0.77* |
| \(\text{SNR}_{\text{pcm}}\)                       | 0.93* |
| LP-based measures:                                |       |
| LP coefficients                                   | 0.06  |
| Reflection coefficients                           | 0.46  |
| Log predictor coefficients                        | 0.11  |
| Log reflection coefficients                       | 0.11  |
| Linear area ratios                                | 0.24  |
| Log-area ratios                                   | 0.62  |
| Itakura distance                                  | 0.59  |
| Linear spectral distance                          | 0.38  |
| Inverse linear spectral distance                  | 0.63  |
| Log spectral distance                             | 0.60  |
| Nonlinear spectral distance                       | 0.61  |
| Frequency variant linear spectral distance         | 0.68  |
| WSSM                                              | 0.74  |
| Composite measures:                               |       |
| Simple and frequency-weighted variant measures     | 0.86  |
| Parametric objective measures                      | 0.82  |

*SNR measures are correlated across only waveform coder distortions. After Queckenhush et al. (1988.)
G. Automatic speech recognition using VQ

Applicability:
For small vocabulary, isolated words recognition, where the spectral contents are sufficiently different among words without temporal constraints.

Method-I: Word-dependent codebooks

Training:
- Collect sufficient data for each word $w_i$, $1 \leq i \leq N$.
For example, for speaker-independent digit recognition, 100 utterances from 100 speakers are often used for each digit, the male and female talkers are balanced.
- For each word $w_i$, train a size-$M$ codebook

$$C^{(i)} = \left\{ y_1^{(i)}, y_2^{(i)}, \ldots, y_M^{(i)} \right\}$$
Recognition:

For the spectral sequence of an unknown utterance $X = \{x_1, x_2, \ldots, x_T\}$:

- Compute the minimum distortion of $X$ w.r.t. each codebook $C^{(i)}$:

\[
D^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \leq j \leq M} d(x_t, y_j^{(i)}) \quad 1 \leq i \leq N
\]

- Classify the utterance $X$ based on minimum distortion:

\[
i^* = \arg \min_{1 \leq i \leq N} D^{(i)}
\]

- Label $X$ by $w_{i^*}$.
Example:

A VQ-based recognition system was built for highly nonconfusable vocabulary of 20 words.

Recognition accuracy:

speaker-dependent: 99%

speaker-independent: 88%

Limitations:

Cannot distinguish between words that can be changed into each other by shuffling temporal ordering of spectral vectors.

Example:

god (/g/ /aa/ /d/) v.s. dog (/d/ /aa/ /g/)

say (/s/ /ey/) v.s. ace (/ey/ /s/)
Method-II: Global codebook

Training:
- Pool together the training data to train a single codebook

\[ C = \{y_1, y_2, \cdots, y_K\} \]

In general, the size of the global codebook is much larger than the size of word-dependent codebooks.
- Estimate the codeword probabilities \( P(y_j | w_i) \):

\[
\hat{P}(y_j | w_i) = \frac{\text{number of times } y_j \text{ occurred in } w_i}{\text{number of training vectors in } w_i}
\]

\[ 1 \leq j \leq K, \ 1 \leq i \leq N \]

Recognition:
- Vector quantize \( X = \{x_1, x_2, \cdots, x_T\} \) into the sequence of codeword indices \( \{m_1, m_2, \cdots, m_T\} \)
- For each word $w_i$, compute the probability of observing the codeword sequence $\hat{X} = \{y_{m_1}, y_{m_2}, \ldots, y_{m_T}\}$:

$$
P(\hat{X}|w_i) = \prod_{t=1}^{T} P(y_{m_t}|w_i)
$$

- Classify the utterance $X$ based on maximum $a$ posteriori probability

$$
i^* = \arg \max_{1 \leq i \leq N} P(w_i|\hat{X})
$$

$$
= \arg \max_{1 \leq i \leq N} \frac{P(\hat{X}|w_i)P(w_i)}{P(\hat{X})}
$$

$$
= \arg \max_{1 \leq i \leq N} P(\hat{X}|w_i)P(w_i)
$$

$$
= \arg \max_{1 \leq i \leq N} P(\hat{X}|w_i) \text{ if } P(w_i) \text{ is uniform}
$$

- Label $X$ by $w_{i^*}$.
In implementation, log probabilities are used in decision rule:

\[ i^* = \arg \max_{1 \leq i \leq N} \log P(\hat{X}|w_i) \]

\[ = \arg \max_{1 \leq i \leq N} \sum_{t=1}^{T} \log P(y_{m,t}|w_i) \]

- Advantage method-II:
  allow sharing of data cross words in deriving the codebook

- Disadvantage of method-II:
  the word-dependent probability terms cannot be reliably estimated from a small amount of training data, causing decision error.

- Method-II is often embedded in subword-unit based speech recognition systems.
Figure 4.20  Effects of cepstral liftering on a log LPC spectrum, as a function of the lifter length ($L = 8$ to 16) (after Juang et al. [11]).
Comparison of correlations between objective measure and subjective speech quality:

| Objective Quality Measure                                      | $|\rho|\,$ |
|-----------------------------------------------------------------|-------|
| SNR                                                            | 0.24* |
| $SNR_{_{\mbox{agg}}}$                                          | 0.77* |
| $SNR_{_{\mbox{tot}}}$                                          | 0.93* |
| **LP-based measures:**                                          |       |
| LP coefficients                                                | 0.06  |
| Reflection coefficients                                        | 0.46  |
| Log-predictor coefficients                                     | 0.11  |
| Log reflection coefficients                                     | 0.11  |
| Linear area ratios                                             | 0.24  |
| Log area ratios                                                | 0.62  |
| Hakuru distance                                                | 0.59  |
| Linear spectral distance                                       | 0.38  |
| Inverse linear spectral distance                               | 0.63  |
| Log spectral distance                                          | 0.60  |
| Nonlinear spectral distance                                    | 0.61  |
| Frequency variant linear spectral distance                     | 0.68  |
| WSSM                                                           | 0.74  |
| **Composite measures:**                                        |       |
| Simple and frequency-weighted variant measures                 | 0.86  |
| Parametric objective measures                                  | 0.82  |

*SNR measures are correlated across only waveform outlier distortions. After Quackenbush et al. (1988.)