ABSTRACT:
Most current speaker recognition systems use Mel frequency cepstral coefficients (MFCC) as the speaker discriminating features. MFCCs are typically obtained using a non-uniform filter bank which emphasizes the low frequency region of the speech spectrum. However, some recent studies have suggested that middle and higher frequency regions of the speech spectrum carry more speaker-specific information. In this work, a general method to obtain cepstral coefficients based on different warped frequency scales is proposed. This method is applied to experimentally investigate the relative importance of specific spectral regions in speaker recognition from vowel sounds.

INTRODUCTION:
Speaker recognition is the task of recognizing a person from his or her voice. A speaker recognition system has three basic functional modules (a) feature extraction, (b) speaker modeling and (c) pattern matching and decision-making. Features derived during training from the speaker’s speech are used to model the speaker. The most popular feature set has been the vector of Mel frequency cepstral coefficients (MFCC) traditionally used also in speech recognition. MFCCs are cepstral coefficients computed on a warped frequency scale based on known human auditory perception. Speaker recognition systems seem to have carried over the legacy of speech recognition in terms of the choice of features. Although it is widely acknowledged that speech recognition and speaker recognition are complementary activities, practically the same features are used for both. Humans can identify a speaker even when listening to unidentifiable utterances. As a matter of fact, speech content and speaker identification are known to be processed in different areas of the human brain. Speech comprehension is based in the left hemisphere of the brain and the right hemisphere is implicated in speaker identification. This suggests different mechanisms for the two functionalities. A study by Sambur to determine signal features that are most effective for speaker recognition, it was found that vowel formants (F2, F3, F4), F2 in nasals and the average pitch were the most effective features for speaker recognition. MFCCs are typically computed by using a bank
of triangular-shaped filters, with the center frequency of the filter spaced linearly for frequencies less than 1000 Hz and logarithmically above 1000 Hz. The bandwidth of each filter is determined by the center frequencies of the two adjacent filters and is dependent on the frequency range of the filter bank and number of filters chosen for design. But for the human auditory system it is estimated that the filters have a bandwidth that is related to the center frequency of the filter. Further it has been shown that there is no evidence of two regions (linear and logarithmic) in the experimentally determined Mel frequency scale. Recent studies on the effectiveness of different frequency regions of the speech spectrum for speaker recognition [3], brought out the importance of frequency regions 0-500 Hz and 3500-4500 Hz. Also, higher frequencies were important for female speakers. A new filter bank front-end was proposed for speaker identification, giving more importance by way of narrower bandwidths to the frequency regions 0 to 1 kHz and 3 kHz to 4.5 kHz. This provided better performance than standard Mel scale filter bank. An alternate to the filter bank method to achieve frequency-scale warping is via the bilinear transform. By suitably fixing a “warping factor” for the given sampling frequency, it is possible to obtain a desired warping function. Thus the bilinear transform provides us a flexible method to achieve a range of warping functions. We use this framework to carry out an experimental study toward determining the optimal choice of frequency-scale warping for the speaker recognition task. The present study is confined to speech vowels, which phonemes are known to contribute the most towards the speaker recognition. Further, since it has been noted in recent literature that the optimal filter bank for speaker discrimination may actually be phoneme-dependent, we present results for each of a set of selected vowels.

The acoustic speech signal contains different kind of information about speaker. This includes “high-level” properties such as dialect, context, speaking style, emotional state of speaker and many others. A great amount of work has been already done in trying to develop identification algorithms based on the methods used by humans to identify speaker. But these efforts are mostly impractical because of their complexity and difficulty in measuring the speaker discriminative properties used by humans. More useful approach is based on the “low-level” properties of the speech signal such as pitch (fundamental frequency of the vocal cord vibrations), intensity, formant frequencies and their bandwidths, spectral correlations, short-time spectrum and others.
PROJECT OVERVIEW:

The Human Vocal apparatus:

Waveform shaping:
• The output of the excitation source is quasi periodic puffs of air.
• These pulses are pressed through the cavities of the throat, mouth and nose to generate the desired sounds.

• To produce speech, air from the lungs is pressed through the throat and released through the mouth and/or nasal cavities.
• The consequent sounds depend on the flow of air and on the shape of these cavities.
• Broadly we may divide overall system into two parts
  • Source – Sub-glottal system
    – Lungs and diaphragm
  • System – Vocal Tract (supra laryngeal cavity)
    – Pharynx, oral and nasal cavities.

Regions in the vocal apparatus:
• **Alveolar ridge** – A short distance behind the upper teeth is a change in the angle of the roof of the mouth. (contains the hard and soft palates).
• **Tongue body/dorsum** – The main part of the tongue, lying below the hard and soft palate.
• **Pharynx** – The cavity between the root of the tongue and the walls of the upper throat.
• **Epiglottis** – The fold of tissue below the root of the tongue. The epiglottis helps cover the larynx during swallowing, making sure (usually!) that food goes into the stomach and not the lungs.

---

**Classes of Speech Sounds**

- **Vowels**
  - iy
  - ih
  - eh
  - ae
  - aa
  - er
  - ax
  - ao
  - uw
  - uh
  - ow

- **Semitones**
  - ay
  - oy
  - aw
  - ey

- **Consonants**
  - w
  - l
  - r
  - y

  - Nasals
    - m
    - n
    - ng

  - Stops
    - b
    - d
    - g
    - p
    - t
    - k

  - Fricatives
    - v
    - dh
    - z
    - zh
    - f
    - th
    - s
    - sh

  - Affricates
    - jh
    - ch

---

**BLOCK DIAGRAM:**

- **PRE-EMPHASIS** → **TIME WINDOW** → **DFT** → **ABSOLUTE** → **FILTER BANK**
- **FEATURE VECTOR** → **DCT** → **LOG**
**DESCRIPTION:**

We will use data base to store different speakers speech (.wav files) using microphone. The frequency-warped cepstral coefficients of 16 kHz sampled speech signal are computed. After pre-emphasis and windowing with a Hamming window of length 20 ms, a high-resolution DFT is used to locate the harmonic peaks and determine the spectral amplitudes at the peaks. The harmonic frequencies are frequency warped using the bilinear transformation. A uniform frequency-interval interpolated spectrum (of I024 points in 0-8 kHz) is calculated from the above warped envelope spectrum.

The cepstral coefficient vector is calculated from the frequency-warped spectrum using Discrete Cosine Transform $c \text{DCT}(\text{Warped Amplitude Spectrum})$ $10 \sim = \log$ The above cepstral vector is used as the feature vector (after discarding the zeroth coefficient since it represents only the signal energy). As a measure of the effectiveness of a given feature for recognition the “divergence” is used in the study. It is a measure of distance or dissimilarity between two classes based upon information theory, and provides means of feature ranking and evaluation of class discrimination effectiveness. A speaker identification test was conducted using vector quantisation as the speaker model. Three utterances of each of the 3 vowels formed the training set for each speaker. The remaining 3 vowel utterances were combined to form the test sentence for each speaker. An eight code-vector VQ with various selected dimensions was used to evaluate each warping function. The average of first differences between the distortion-distance of the first and second matched speaker (in case of successful recognition) was used as a measure of discrimination between two speaker classes. (All the warping considered resulted in 100% recognition, except for one test sentence in case of 12D Mel warping).

Figure 4 shows the result of this measure. Ozgur warping seems to give the best discrimination. Also it is noted that linear warping gives better discrimination than mel and Bark warping in contradiction to divergence measure. This may be due to the fact that here we are comparing only the first two matched speakers rather than averaging the divergences between speakers. Also in divergence measure we are using the entire data for each vowel, while we are clustering in VQ and only using small portion of the original data for testing.
SOFTWARE USED:
MATLAB:

MATLAB (matrix laboratory) is a numerical computing environment and fourth-generation programming language. Developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and Fortran.

Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. An additional package, Simulink, adds graphical multi-domain simulation and Model-Based Design for dynamic and embedded systems.

In 2004, MATLAB had around one million users across industry and academia. MATLAB users come from various backgrounds of engineering, science, and economics. MATLAB is widely used in academic and research institutions as well as industrial enterprises.

Applications:

Practical applications for automatic speaker recognition(identification) are obviously various kinds of security systems. Human voice can serve as a key for any security objects, and it is not so easy in general to lose or forget it. Another important property of speech is that it can be transmitted by telephone channel, for example. This provides an ability to automatically identify speakers and provide access to security objects by telephone. Nowadays, this approach begins to be used for telephone credit card.
purchases and bank transactions. Human voice can also be used to prove identity during access to any physical facilities by storing speaker model in a small chip, which can be used as an access tag, and used instead of a pin code. Another important application for speaker identification is to monitor people by their voices. For instance, it is useful in information retrieval by speaker indexing of some recorded debates or news, and then retrieving speech only for interesting speakers. It can also be used to monitor criminals in common places by identifying them by voices. In fact, all these examples are actually examples of real time systems. For any identification system to be useful in practice, the time response, or time spent on the identification should be minimized. Growing size of speaker database is also common fact for practical systems and can also lead to system optimization.

CONCLUSIONS:
This is an experiment conducted with speech signal sampled at 16 kHz. For this signal emphasizing the lower frequency results in improved divergence measure. Also from the better performance of Ozgur warping the importance of frequency around 3 to 5 kHz can be observed. It seems that for speaker recognition there can be better warping than commonly used Mel scale warping. This result may be valid for the individual phonemes in question, and may not hold across other phonemes. Other phonemes are to be studied, also with more speakers.

REFERENCES:


Sites:

“www.ieee.org”
“www.mathworks.com”
“www.wikkipedia.com”
“www.Americanorf.com”