Comparison of Speech Normalization Techniques

David McCarten
E6820 Student, Columbia University
March 9, 2008

1. Goals of the project
2. Reasons for speech normalization
3. Speech normalization techniques
4. Spectral warping
5. Test setup with SPHINX-4 speech recognition system
6. Initial test results

Goals of the project

David McCarten
E6820 Student, Columbia University
March 9, 2008

1. Survey speech normalization techniques.
2. Convey understanding of how these techniques work.
3. Determine importance of normalization to speech recognition improvement.
4. Test speech recognition improvement with speakers of varying dialects, especially ESL speakers.
5. Compare performance of existing techniques.
6. Recommend future work toward two goals:
   A. Accurate conversational speech recognition.
   B. Automatic meeting minutes transcription generation.

REFERENCES:

[1] Li Lee, Richard C. Rose, AT&T Bell Labs, Murray Hill, NJ, USA, "Speaker Normalization Using Efficient Frequency Warping Procedures", Acoustics, Speech, and Signal Processing, 1996, ICASSP-96. Conference Proceedings, Publication Date: 7-10 May 1996

[2] Donald Bailey, Warwick Allen, Serge Demidenko, "Spectral Warping Revisited", Proceedings of the Second IEEE International Workshop on Electronic Design, Test and Applications (DELTA'04), 2004

[3] Sadaoki Furui, "Steps Toward Natural Human-Machine Communication in the 21st Century", Deparment of Computer Science, Tokyo Institute of Technology,Proc. COST249 Workshop,"Voice Operated Telecom Services, 2000

[4] Steven Wegmann, Don McAllaster, Jeremy Ortoff, Barbara Peskin, "Speaker Normalization on Conversational Telephone Speech", Dragon Systems, Inc., Acoustics, Speech, and Signal Processing, 1996. ICASSP- Conference Proceedings., 1996 IEEE International Conference

[5] Alejandro Acero, Richard M. Stern, Dept. of Electrical Engineering and Computer Engineering and School of Computer Science, Carnegie Mellon University, Pittsburgh, Penn, USA, "Robust Speech Recognition by Normalization of the Acoustic Space", Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991, Publication Date: 14-17 Apr 1991

[6] Puming Zhan, Alex Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech recognition", May, 1997

[7] John McDonough, William Byrne, Xiaoqiang Luo, Center for Language and Speech Processing, The john Hopkins University, Baltimore, MD, USA, "Speaker Normalization with All-Pass Transforms", 1998

[8] Charles R. Jankowski Jr., Richard P. Lippmann, MIT Lincoln Laboratory, Lexington, MA, USA, "Comparison of Auditory Models for Robust Speech Recognition", 1998

Reasons for speech normalization

David McCarten
E6820 Student, Columbia University
March 9, 2008

Automatic Adaptation to Voice Variation

1. Acoustic variation in speech
    - A. Physical and psychological condition of the speaker
    - B. Telephone, microphones, network conditions
    - C. Noise
        1. Background noise (additive)
        2. Reverberations
    - D. Other speakers
    - E. Speaking styles
    - F. Distortion, echoes, dropouts
    - **G. Speaker characteristics**
        1. **Pitch**
        2. **Gender**
        3. **Dialect**
    - H. Task/context
        1. Dialogue
        2. Dictation
        3. Interview
    - I. Microphone
        1. Distortion
        2. Electrical noise
        3. Directional characteristics

Reasons for speech normalization

David McCarten
E6820 Student, Columbia University
March 9, 2008

Automatic Adaptation to Voice Variation

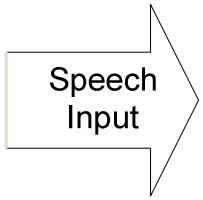
J. Spontaneous speech recognition must deal with variations:
    1. extraneous words
    2. out-of vocabulary words
    3. ungrammatical sentences
    4. disfluency
    5. partial words
    6. repairs
    7. hesitations
    8. repetitions
K. Error rate is 3-4 times greater in SI compared to SD systems

Speech normalization techniques

David McCarten
E6820 Student, Columbia University
March 9, 2008

Speaker adaptation or normalization

1. Types
   A. Supervised
   B. Unsupervised
      1. Unsupervised, instantaneous and incremental speaker adaptation combined with automatic detection of speaker changes is ideal.
2. Must adapt many phonemes using a limited number of utterances.
   1. Ergo, must use adequate modeling of speaker-to-speaker variability.
3. Main method to deal with voice variations



   A. Microphone
      1. Close-talking microphone
      2. Microphone array
   B. Analysis and feature extraction
      1. Auditory models
         a. EIH - Ensemble Interval Histogram
         b. SMC – Speech and Multimedia Communication
         c. PLP – Perceptual Linear Predictive Speech Analysis

Speech normalization techniques

David McCarten
E6820 Student, Columbia University
March 9, 2008

Speaker adaptation or normalization

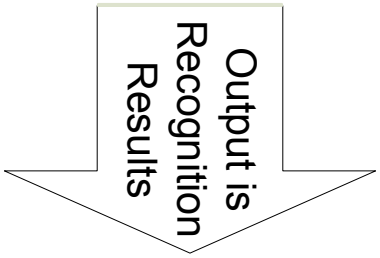
3. Main method to deal with voice variations
    C. Feature-level normalization/adaptation
        1. Adaptive filtering
        2. Noise subtraction
        3. Comb filtering
        4. Spectral mapping
        5. Cepstral mean normalization
        6. Δ cepstra
        7. RASTA
    D. Model-level normalization/adaptation – input is reference templates/models
        1. Noise addition
        2. HMM (de)composition(PMC)
        3. Model transformation (MLLR)
        4. Bayesian adaptive learning
    E. Distance/similarity measures
    F. Frequency weighting measure
    G. Weighted cepstral distance
    H. Cepstrum projection measure

Speech normalization techniques

David McCarten
E6820 Student, Columbia University
March 9, 2008

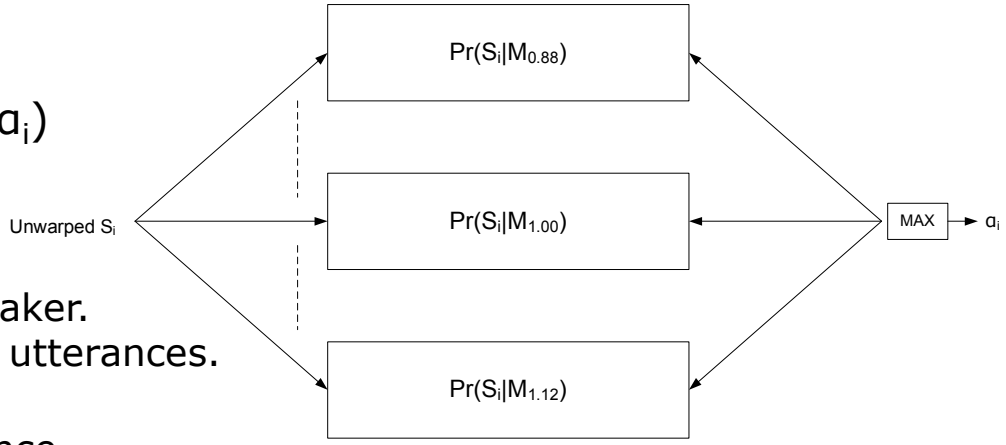
Speaker adaptation or normalization

4. Robust matching
   A. Word spotting
   B. Utterance verification
5. Linguistic processing
   A. Language model adaptation



Spectral warping for a HMM speech recognition system per [1]

David McCarten
E6820 Student, Columbia University
March 9, 2008

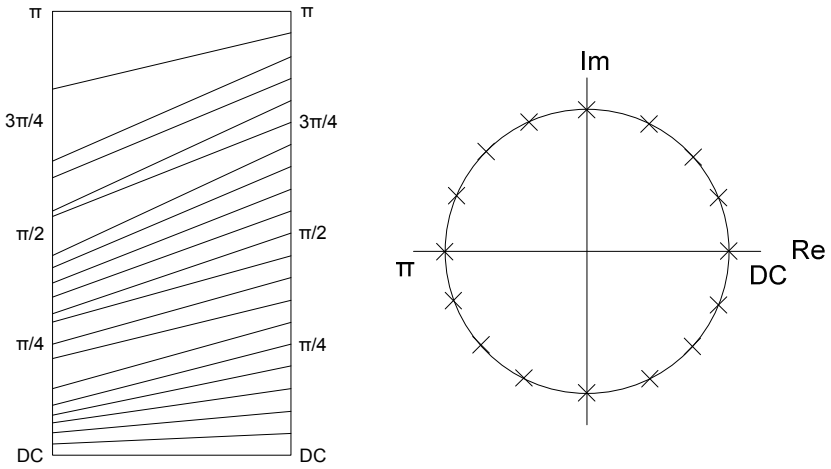


1. Estimate optimal warping factor (α_i)
   A. During HMM training.
   B. Again during HMM recognition.
2. Training
   A. One α_i is determined for each speaker.
   B. One HMM is built from all warped utterances.
3. Recognition
   A. Estimate α_i based on input utterance.
   B. Decode utterance using warped feature vectors, e.g. MFCC.
4. Estimation of α_i
   A. Maximum likelihood with respect to a specific HMM.
   B. Optimal warping factor $\alpha_i = \text{argmax}[\Pr(X_i|\lambda, W_i)]$.
      1. λ : set of HMM models.
      2. X_i: α_i warped cepstrum domain observation vectors for speaker i.
      3. W_i: transcriptions.
   C. Practical range of : 0.88 to 1.12, accounting for the 25% range of vocal tract lengths.

Spectral Normalization

David McCarten
E6820 Student, Columbia University
March 9, 2008

1. Vocal tract length varies from person to person
   A. Formant frequency peaks are inversely proportional to vocal tract length.
   B. Formant center frequencies vary by as much as 25% between speakers.
   C. Frequency warping is a technique to re-map speech around the z-domain unit circle, such that all utterances seem to be generated by the same vocal tract.



Spectral warping using SPHINX-4 HMM speech recognition system

David McCarten
E6820 Student, Columbia University
March 9, 2008

Frequency warping performed in two stages per [5], resulting in 10% error rate reduction.

1. Bilinear transformation performed with $\alpha=0.6$
    A. LPC-cepstrum mapped into pseudo mel scale
2. Another transform applied with a variable warping parameter, $\Delta\alpha$
    A. $\Delta\alpha$ chosen to minimize VQ error.
    B. Maintain average of zero between male and female speakers.

Test Configuration

David McCarten
E6820 Student, Columbia University
March 9, 2008