Interspeech 2016

# Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition

Soo Jin Park[1], Caroline Sigouin[2], Jody Kreiman[3], Patricia Keating[4], Jinxi Guo[1], Gary Yeung[1], Fang-Yu Kuo[1] and Abeer Alwan[1]

[1]Department of Electrical Engineering, [3]Department of Head and Neck Surgery, School of Medicine, [4]Department of Linguistics, University of California Los Angeles, USA
[2]Department of Language, Linguistics and Translation, Université Laval, Québec, Canada

## ABSTRACT

This study used short test utterances (2-3sec) to investigate the effect of within-speaker variability on state-of-the-art ASpR system performance. For 25 female speakers, the short utterances combined with affect mismatch degraded system performance by 106%.

Considering that humans are more robust to within-speaker variability, human perception experiments were also conducted to understand how humans perceive speaker identity. In this study, a model is proposed to predict the perceptual dissimilarity between tokens.

Results showed that a set of voice quality features provides information that complements MFCCs. By fusing the feature set with MFCCs, human response prediction RMS error was reduced by 12% compared to using MFCCs alone. In ASpR experiments with short utterances from 50 speakers, the voice quality feature set decreased the error rate by 11% when fused with MFCCs.

## CONTACT

**Park, Soo Jin**
sj.park@ucla.edu
PhD Candidate

Speech Processing and Auditory Perception Lab.,
University of California, Los Angeles

http://www.seas.ucla.edu/spapl

## INTRODUCTION

### Machine vs. Human Speaker Recognition

- Automatic speaker recognition (ASpR)
  - Remarkable improvements with i-vector framework
  - Performance still degrades when utterances are short
  - Within-speaker variability also degrades the performance e.g. emotional speech [1]
- Human speaker recognition
  - Able to distinguish speakers with high accuracy even from very short utterances
  - Perform better with within-speaker variability than machines

### Motivations

- Obtaining insights into how human recognize speakers may improve ASpR systems
- Predicting perceived speaker identity itself is an interesting topic e.g. Forensics

### Features for Speaker Recognition

- Features for human speaker recognition
  - No single set of acoustic parameters associated with human speaker recognition has been identified
  - Humans recognize voices as complex, integral auditory patterns [2]
- Mel-frequency cepstral coefficients (MFCCs)
  - Most popular in ASpR applications
  - Represent vocal tract information well, but not the voice source
- Voice source information in ASpR
  - Studies showed the effectiveness
    - Espy-Wilson et al. used acoustic parameters consisting of both voice source and vocal tract features [3]
    - Mazaira et al. used cepstral coefficients from the inverse-filtered signal [4]
  - Still has not been utilized extensively
- In this study, voice source features are added to other acoustic features to better represent voice quality and speaker identity

### Objectives

1. To find out if voice quality features are useful in modeling human judgements of speaker identity
2. To study how to use voice quality features to improve ASpR systems when there is variability and the utterances are short

## DATABASE

### UCLA Database

- To study both within- and between-speaker variability
- Multiple tasks per speaker
  - Sustained vowels, read sentences, instructions, affective speech, conversational speech, and exaggerated prosody
  - Tasks per session are summarized in Table 1.
- Large number of speakers
  - More than 100 female and 100 male speakers
  - UCLA undergraduate students
- High quality recording
  - Sound-attenuated booth
  - ½" Brüel & Kjær microphone
  - Sampling rate of 22kHz

**Table 1.** Speech tasks in UCLA database

| Session | A | B | C |
|---|---|---|---|
| **Sustained vowel /a/** | 3 repetitions | | |
| **Read sentences** | 2 repetitions of 5 Harvard sentences | | |
| **Instructions** | 30-sec | N/A | N/A |
| **Experience telling** | neutral (30-sec) | happy (30-sec) | annoyed (30-sec) |
| **Conversational speech** | N/A | phone-call (2-min) | N/A |
| **Exaggerated prosody** | N/A | N/A | pet-directed (1-min) |

## VOICE QUALITY FEATURES

### Voice Quality Feature (VQual) Set

- Voice quality: A perceptual response to an acoustic voice signal
- Measured using a psychoacoustic model [5]
- F0, F1, F2, F3, H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k, and cepstral peak prominence (CPP)
  - Hn indicates the amplitude of n-th harmonic component (see Figure 1)
  - H2k and H5k indicate the amplitude of the harmonic components near 2kHz and 5kHz respectively
  - The asterisks (*) indicate that the effect of formants is corrected
- Selected based on a study to find the necessary and sufficient set of features contributing to perceived voice quality [5, 6]
- In a previous study, VQuals predicted listeners' confusion reasonably well from sustained vowel /a/ sounds [7]

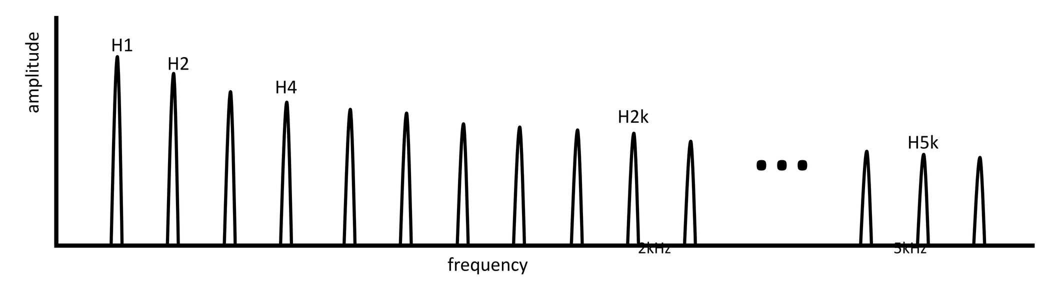



**Figure 1.** A schematic for the source spectral model for the voice quality feature set.

## HUMAN PERCEPTION EXPERIMENTS

### Method

- Stimuli
  - Pairs of read sentences
    - Two repetitions of 2 different sentences from 2 sessions
      - » "A pot of tea helps to pass the evening"
      - » "The soft cushion broke the man's fall"
    - Three female speakers (8x3 = 24 utterances)
    - 30 same-speaker pairs and 48 different-speaker pairs
- Listeners
  - 15 normal-hearing UCLA students and staff members
  - Judged whether each pair represents one speaker or two different speakers
  - Self-paced

### Results

- Highly accurate even when the utterances were short (< 3sec)
- More accurate on read sentences than on isolated vowels
  - Sentences: 89% accurate
  - Vowels: 69% accurate [7]

## MODELING HUMAN RESPONSES

### Method

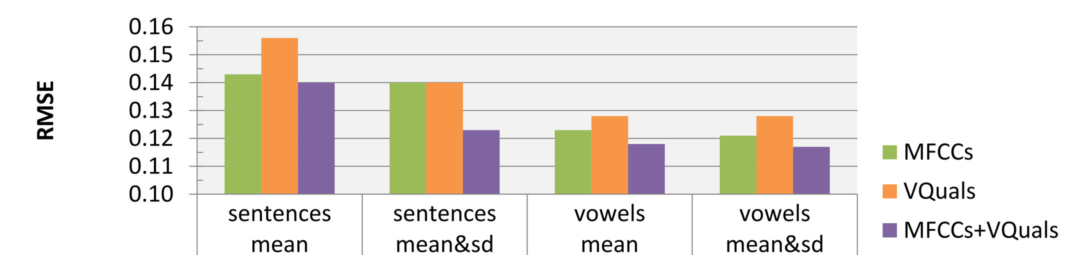
- Dissimilarity score

$$d = \begin{cases} u, & \text{if 'same speaker' response} \\ 11 - u, & \text{if 'different speaker' response} \end{cases}$$

  - Averaged dissimilarities $\bar{d}$ ranged from 0 to 10
  - Zero dissimilarity was assigned to identical token pairs
- Token distance in a perceptual space
  - Multi-dimensional scaling (MDS, [8])
    - 6-dimensional non-metric MDS
    - Euclidean distance between token pairs of all possible combination
- Acoustic features
  - Baseline: 20-MFCCs + Δ + ΔΔ
  - Voice quality features (VQuals) + Δ + ΔΔ
- Perceptual dissimilarity prediction method
  - Linear regression
    - Target: Euclidean distance between two tokens in the MDS perceptual space
    - Predictors: differences in means only, or differences in means and standard deviations of the features between the two tokens

### Results and Discussion

- Performance measure
  - Root-mean-squared errors (RMSE) are shown in Figure 2
- Discussion
  - VQuals provided complementary information to MFCCs
  - Acoustic features did less well at predicting human responses for the sentences than for the vowels
  - Score-level fusion was also tried, but no improvement was found over concatenating the features
  - Since voice quality features provided complementary information to MFCCs, they might also improve automatic speaker recognition systems




| | sentences mean | sentences mean&sd | vowels mean | vowels mean&sd |
|---|---|---|---|---|
| MFCCs | 0.143 | 0.140 | 0.123 | 0.121 |
| VQuals | 0.156 | 0.140 | 0.128 | 0.128 |
| MFCCs+VQuals | 0.140 | 0.123 | 0.118 | 0.117 |

**Figure 2.** Perceptual dissimilarity prediction performance in RMSE using either MFCCs, VQuals, or the combination of the two. The prediction target was the Euclidean distance in the perceptual space, and the predictors were either the differences in means of the features, or the differences in both means and standard deviations.

## AUTOMATIC SPEAKER RECOGNITION

### Analyzing the Effect of Within-Speaker Variability

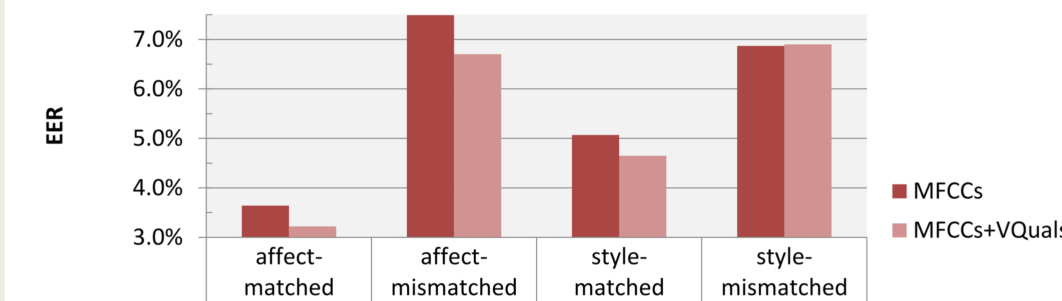
- Stimuli
  - Speakers: 25 female, 25 male speakers
  - Read sentences and affective speech
- Experiment conditions
  - Matched condition
    - Enrolling the speakers with data containing variability and testing with the known variability
  - Mismatched condition
    - Testing with the utterances with unseen variability
  - Variability of interest
    - Session (read sentences in session A, B, and C)
    - Affect (neutral/happy/annoyed)
    - Speaking-style (read/spontaneous)

### Standard Automatic Speaker Recognition (ASpR) System Performance

- System Setup
  - Features: 20-MFCCs + Δ + ΔΔ
  - i-vector/PLDA speaker verification system with the Kaldi toolkit [9]
  - Gender-dependent
- Performance Measure
  - Equal error rates (EER) are shown in Figure 3
- Results and Discussion
  - Session variability did not affect system performance
  - Affect variability and speaking-style variability caused a notable degradation
  - This result might be dependent on the lexical content
  - Further analysis is needed with more speakers and more controls

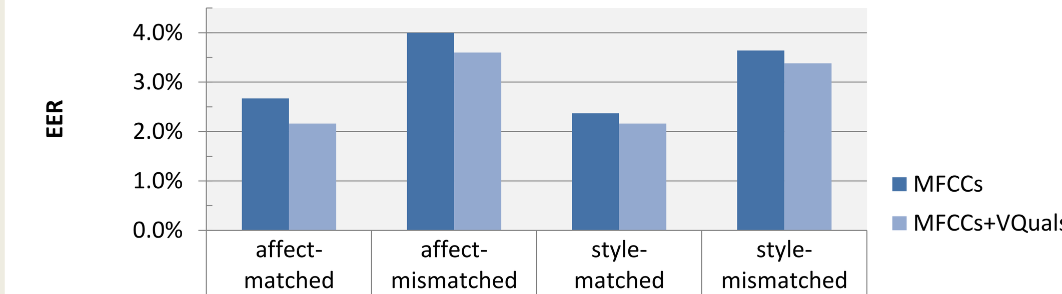
### Voice Quality Feature Effect on the ASpR System

- System Setup
  - Score-level fusion between the baseline system and a system using VQual features + Δ + ΔΔ
- Results and Discussion
  - Fusing VQual features improved the system performance, providing complementary information to MFCCs
    - Affect-matched condition: 11.60% relative improvement for female, 19.11% for male speakers
    - Affect-mismatched condition: 10.49% for female, 9.89% for male speakers
    - Style-mismatched condition: 8.26% for female, 8.78% for male speakers
  - A difference in EER between matched and mismatched conditions remained
  - Voice quality may be varying significantly according to the emotional status and speaking-style of the speaker




| | affect-matched | affect-mismatched | style-matched | style-mismatched |
|---|---|---|---|---|
| MFCCs | 3.64% | 7.49% | 5.07% | 6.87% |
| MFCCs+VQuals | 3.22% | 6.70% | 4.65% | 6.90% |

**(a) System performance for female speakers**




| | affect-matched | affect-mismatched | style-matched | style-mismatched |
|---|---|---|---|---|
| MFCCs | 2.67% | 4.00% | 2.37% | 3.64% |
| MFCCs+VQuals | 2.16% | 3.60% | 2.16% | 3.38% |

**(b) System performance for male speakers**

**Figure 3.** Equal error rate (EER) for the automatic speaker recognition systems for four different conditions (affect-matched/mismatched and style-matched/mismatched). The baseline system (MFCCs) uses only MFCC features while the proposed system (MFCCs+VQuals) fuses the score from the baseline system and the system using voice quality features. The performance for (a) female and (b) male speakers are depicted in separated panels.

## CONCLUSION

### Voice Quality Features (VQuals)

- Used to better represent speaker identity

### Modeling Human Responses

- VQual features provided complementary information to MFCCs
- RMSE decreased as much as 11.80% for read sentences

### Automatic Speaker Recognition

- State-of-the-art ASpR system performed worse when there was within-speaker variability and the utterances were short
- VQuals did not necessarily improve the robustness to within-speaker variability, but did improve ASpR system performance in most conditions

### Further Studies

- Improving algorithm to extract VQual features
- Including perception experiments with more speakers and more variability to reveal how robust humans are to a wide range of variabilities

## REFERENCES

[1] T. Wu et al., "MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition." In *IEEE Odyssey 2006: Workshop on Speaker and Language Recognition*, 2006.

[2] J. Kreiman and D. Sidtis, *Foundations of voice studies*. Wiley-Blackwell, 2011.

[3] C. Espy-Wilson et al., "A new set of features for text-independent speaker identification," in *Proc. of ICSLP*, pp. 1475-1478, 2006.

[4] L. Mazaira-Fernandez et al., "Improving speaker recognition by biometric voice deconstruction," *Frontiers in Bioengineering and Biotechnology*, vol.3, 2015.

[5] M. Garellek et al., "Modeling the voice source in terms of spectral slopes," J. Acoust. Soc. Am., vol. 139, pp. 1404-1410, 2016.

[6] M. Garellek et al., "Perceptual sensitivity to a model of the source spectrum," *Proc. of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[7] J. Kreiman et al., "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Interspeech*, 2015.

[8] J. Kruskal and M. Wish, *Multidimensional Scaling*. Sage, 1978.

[9] D. Povey et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com