Interspeech 2017

Using Voice Quality Features to Improve Short-Utterance Text-Independent Speaker Verification Systems

Soo Jin Park[1], Gary Yeung[1], Jody Kreiman[2], Patricia A. Keating[3], and Abeer Alwan[1]

[1]Dept. of Electrical Engineering, [2]Dept. of Head and Neck Surgery, School of Medicine, [3]Dept. of Linguistics, University of California Los Angeles, USA

Introduction

❖Within-Speaker Variability

- A speaker's voice varies by phonetic content, speaking style, etc.
- This variability is a challenge for automatic speaker verification (ASV)
  - Especially true when utterances are short

❖Proposed voice quality (VQual) features

- Humans utilize voice attributes to recognize speakers (Schweinberger et al., 2014)
- Human performance does not degrade much by limited phonetic content and utterance length
- Thus, voice quality might also provide important information for short-utterance text-independent ASV

❖Previous Studies

- Kreiman et al. (2014) and Garellek et al. (2016) introduced a psycho-acoustic model of voice quality which accounts for listeners' judgement of the quality of sustained vowel sounds
- Park et al. (2016) applied voice quality features to ASV
  - The features were complementary to widely-used MFCC features
  - Features were effective in predicting human speaker perception and improved ASV performance

❖Objective of the Current Study

1. Evaluate voice quality features on general ASV tasks including very short utterances (2 sec)
2. Improve the voice quality feature set to better represent speaker identity in ASV tasks

Database

❖UCLA Speaker Variability Database

- Includes both within- and between-speaker variability
- Multiple tasks per speaker (Table 1)
- Large number of speakers (all UCLA undergraduate students)
  - More than 100 female and 100 male speakers
- High quality recording
  - Sound-attenuated booth, ½" Brüel & Kjær microphone
  - Sampling rate of 22kHz

Table 1. Speech tasks in the UCLA SV 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)

Feature Selection

❖Voice Quality Feature (VQual) Sets

- **VQual1** (Park et al., 2016)
  - F0, F1, F2, F3, H1*-H2*, H2*-H4*, H4*-H2k*, H2k*-H5k, and cepstral peak prominence (CPP)
    - Hn: the amplitude of n-th harmonic
    - H2k, H5k: the amplitude of the harmonic near 2kHz and 5kHz
    - Asterisks (*): the effect of formants is corrected
- **Explored the following:**
  - The effectiveness of the correction formula when a formant is very close to a harmonic
  - The role of formant amplitudes, H1*-A1*, H1*-A2, and H1*-A3*, since they may also represent voice quality (Hanson, 1997). Note that An represents the amplitude of the n-th formant

❖Method

- Improve voice quality feature set to better represent speaker identity in ASV tasks
- Feature Relevance Measure
  - Features with large between-speaker variability and small within-speaker variability are desirable using the F-ratio
  - F-ratio (Nicholson et al., 1997)

$$F = \frac{\text{between speaker variance}}{\text{within speaker variance}} = \frac{1/M \sum_{i=1}^{M}(\mu_i - \mu)^2}{1/M \sum_{i=1}^{M} \sigma_i^2}$$

    - M: the number of speakers; μ_i, μ : the within-speaker mean and global mean, respectively; σ_i^2 : the within-speaker variance
- Analysis Conditions
  - Content variability conditions
    - Mixed phonetic content: 5 subsets of randomly chosen read sentences (the F-ratios were computed within the subsets and averaged)
    - Separate phonetic content: 5 subsets of same sentences
  - Speaking style variability conditions
    - Mixed speaking style: 2 subsets of randomly chosen read sentences and pet-directed speech segments
    - Separate speaking style: read-only or pet-directed only

❖Results and Discussion

- The features with high F-ratios were similar between the content and style variability cases
  - Partially because content variability is also present in the style variability subsets
- F-ratios in the pet-directed subset were high
  - Each speaker had a unique style of talking to pets
- Feature set modification
  - The Hn-Hm features had higher F-ratios without formant correction than with it
  - The An features had the highest F-ratios among the features using the formant amplitudes
  - **Hence, a new set (VQual2) is proposed**: F0, F1, F2, F3, H1-H2, H2-H4, H4-H2k, H2k-H5k, A1, A2, A3 and CPP

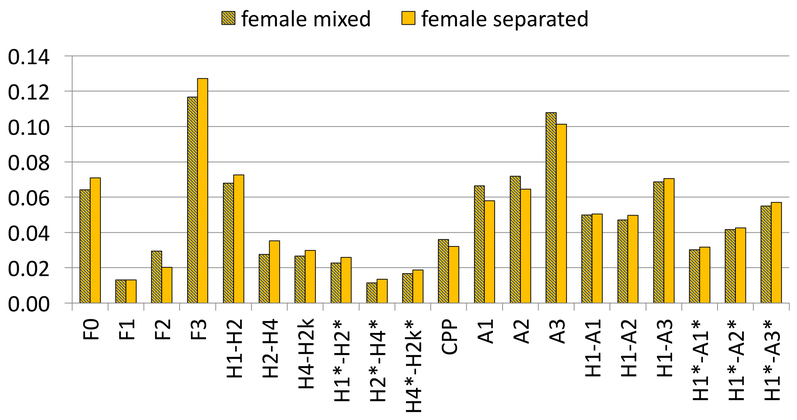


Figure 1 (a). Computed F-ratios of various voice quality features for content variability from female voices. F-ratios from male voices are not shown because they had similar tendency to those of females.

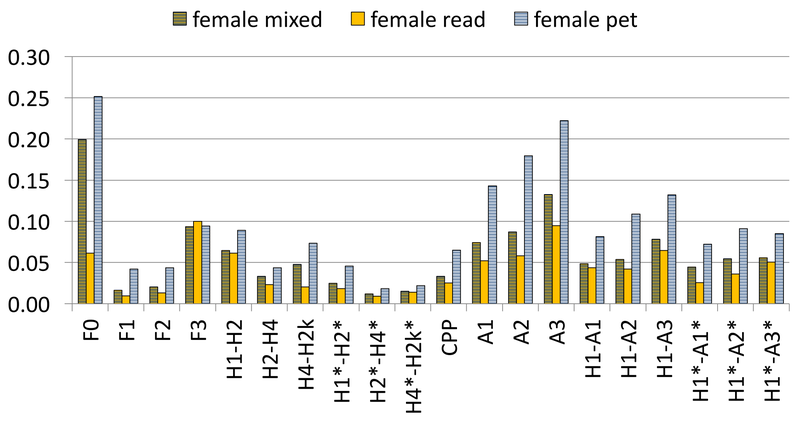


Figure 1 (b). Computed F-ratios of various voice quality features for style variability from female voices.

ASV Speaker Variability Analysis

❖Method

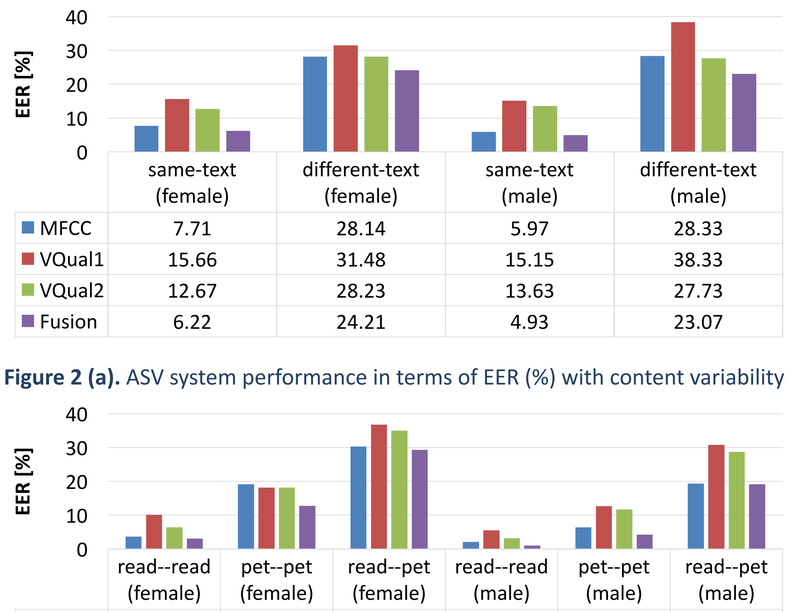
- System Setup
  - i-vector/PLDA system
  - Feature sets
    - 20-MFCCs + Δ (baseline)
    - VQual1 + Δ + ΔΔ
    - VQual2 + Δ + ΔΔ
  - Linear PLDA score fusion between MFCC and VQual2

$$s = \alpha s_{\text{VQual}} + (1-\alpha) s_{\text{MFCC}}$$

    - α : a parameter in a range between 0 and 1; s_{VQual} ,s_{MFCC} : the PLDA score from VQual2 and MFCCs, respectively
- Evaluation Data
  - Read sentences and pet-directed speech in UCLA SV Database
  - 100 female and 100 male speakers
- Analysis Conditions
  - Content variability conditions
    - Same text: Enrollment and testing used different tokens of the same sentences
    - Different text: Enrollment and testing used different sentences
  - Style variability conditions
    - Same style: Read or pet-directed speech for both enrollment and testing
    - Different style: Read sentences for the enrollment and pet-directed speech for testing and vice versa

❖Results and Discussion

- Using only MFCCs resulted in a relative error increase of at least: 265% in content-mismatched conditions and 730% in style-mismatched conditions possibly due to:
  - Feature distortion by the exaggerated prosody
  - Limited phonetic content in the pet-directed speech samples
- VQual2
  - Improved performance over VQual1 in all conditions
  - Exceeded or matched MFCC performance in some conditions
- Score fusion (MFCC+VQual2)
  - Decent improvements in all conditions
  - Notable 33% improvement for the pet—pet trials
  - Little improvement for the style-mismatched condition
    - VQual2 features appear to be affected by speaking style



	same-text (female)	different-text (female)	same-text (male)	different-text (male)
MFCC	7.71	28.14	5.97	28.33
VQual1	15.66	31.48	15.15	38.33
VQual2	12.67	28.23	13.63	27.73
Fusion	6.22	24.21	4.93	23.07

Figure 2 (a). ASV system performance in terms of EER (%) with content variability

	read--read (female)	pet--pet (female)	read--pet (female)	read--read (male)	pet--pet (male)	read--pet (male)
MFCC	3.65	19.19	30.3	2.13	6.38	19.3
VQual1	10.1	18.18	36.77	5.48	12.72	30.85
VQual2	6.4	18.18	35.06	3.19	11.7	28.72
Fusion	3.03	12.79	29.29	1.06	4.25	19.15

Figure 2 (b). ASV system performance in terms of EER (%) with style variability

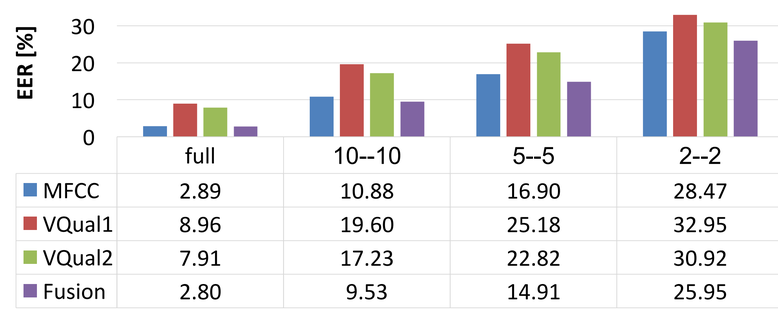
ASV Performance with the NIST Database

❖Method

- System Setup: Same system as in the previous section
- Evaluation Data
  - NIST SRE10 condition 5 extended tasks
  - Full utterances (5 min) and speech segments of the same utterances cut into 10, 5, and 2 seconds

❖Results and Discussion

- VQual2
  - Around 2% absolute improvement from VQual1 in all conditions
- Score fusion (MFCC+VQual2)
  - A relative improvement of around 12% for 10—10 sec and 5—5 sec trials compared to using only MFCCs
  - Small improvement for full (5 min), and 9% improvement for 2—2 sec



	full	10--10	5--5	2--2
MFCC	2.89	10.88	16.90	28.47
VQual1	8.96	19.60	25.18	32.95
VQual2	7.91	17.23	22.82	30.92
Fusion	2.80	9.53	14.91	25.95

Figure 3. ASV system performance in terms of EER (%) with the NIST SRE10 database

Conclusion

❖Voice Quality Feature (VQual) Set Improvement

- Feature set was improved using the F-ratio criterion

❖ASV Speaker Variability Analysis

- Content/style mismatch between the enrollment and test utterances causes a significant error increase
- Score fusion of VQual2 features with MFCCs improved content mismatch performances
- The fusion did not improve performance much for style mismatched conditions
- The performance gain in pet—pet trials suggests VQuals' ability to capture speakers' idiosyncratic ways of exaggerating prosody

❖ASV Short-Utterance Evaluation

- Only short utterances benefit from using VQual2 features

❖Further Studies

- Additional features, e.g. prosodic and subglottal features
- Comparison between human and machine speaker recognition when within-speaker variability is large

References

Garellek, M. et al., 2016, "Modeling the voice source in terms of spectral slopes," J. Acoust. Soc. Am., vol. 139.

Hanson, H. M., 1997, "Glottal Characteristics of Female Speakers: Acoustic Correlates," J. Acoust. Sco. Am., vol. 101, no. 1.

Kreiman, J. et al, 2014, "Toward a Unified Theory of Voice Production and Perception," Loquens, vol. 1, no. 1.

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

Nicholson, S. et al., 1997, "Evaluating Feature Set Performance using the F-ratio and J-measures," in Proc. *Eurospeech*.

Park, S. J. et al., 2016, "Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition," in Proc. *Interspeech 2016*.

Schweinberger, S.R. et al., 2014, "Speaker Perception," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5 no. 1.