PERCEPTUAL RELEVANCE OF OBJECTIVELY MEASURED DESCRIPTORS FOR SPEAKER CHARACTERIZATION

Burhan F. Necioğlu[‡] *Mark A. Clements*[‡] *Thomas P. Barnwell III*[‡] *Astrid Schmidt-Nielsen*[‡‡]

[‡]Center for Signal and Image Processing
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

[‡‡]Code 5513, Naval Research Laboratory
4555 Overlook Ave SW
Washington, DC 20375, USA

ABSTRACT

Subjective testing of speaker recognizability is an intricate, time consuming and very expensive process, but using objectively measurable descriptors to augment the subjective speaker recognizability tests could result in increased efficiency and reliability. This paper describes our investigation into the relevancy of a set of objective descriptors to human perception of speaker identity through multidimensional scaling (MDS) of subjective speaker pair similarity judgments. The evaluated objective descriptors can achieve same/different detection error rates as low as 4.13% for male speaker pairs, and 8.17% for female speaker pairs, with only 3 seconds of speech. Five descriptors related to glottal, vocal tract and prosodic features were found to have significant correlations with the perceptual dimensions of the MDS solutions.

1. INTRODUCTION

Evaluation of speaker recognizability across a communication system, or channel, involves determination of how well the perceived identity of a user is preserved at the receiving end. Naturally, this determination should also address the preservation of distinguishability between the voices of two different users, given that they do not "sound like each other" to start with. Speaker recognizability has long been identified as a component of the evaluation process of communications systems [1], but the intelligibility and voice quality aspects of evaluation have taken relative precedence[2]. However, with more widespread use of lower bit rate speech coders, speaker recognizability emerges as an additional major issue. Still, subjective testing of speaker recognizability, as any other subjective tests, is intricate, time consuming and very expensive, so potentially, using objectively measurable descriptors to augment the subjective speaker recognizability tests could result in increased efficiency and reliability.

Previously, we reported on the speaker discrimination merit and reliability of some objectively measurable descriptors[3, 4]. This paper describes our incorporation of subjective speaker dissimilarity judgments to the evaluation process of objectively measurable descriptors, with the goal to determine their relevancy to the human perception of speaker identity.

Voiers pioneered the research for the determination of the underlying structure of perceptual voice characterization by humans, with the motivation of finding a basis for subjective evaluation of speaker recognizability [5, 6]. He found that listeners' ratings of speaker voices depended upon at least eight underlying orthogonal dimensions. Using subjective dissimilarity ratings of utterances from speaker pairs and acoustic measurements in a multidimensional scaling (MDS) context has been another approach followed by other investigators. In these tests the types of speech utterances used range from sustained vowels [7, 8] and four-phoneme monosyllabic words [9] to a single sentence spoken by all the test speakers [10]. Although slightly varying across studies, the acoustic measurements taken from the speech waveforms were generally based upon pitch and formant frequencies, averaged spectra, and duration information, and all required human interaction for the measurement process. Across these studies, correlations between the acoustic measurements and the dimensions of the resultant MDS solutions vary depending upon the speaker and listener set sizes and utterance types, but average pitch emerges as the acoustic measurement with the highest contribution to the perceived speaker identity difference in all these studies.

Past investigations on speaker recognizability and perceptual characterization of speakers had to utilize a relatively limited number of objectively measurable physical/acoustic parameters of the speech waveform due to various reasons. Perhaps, the leading limitation was the lack of sufficient computing power and equipment. Researchers utilized methods and instruments that required significant user input, time and attention. The goal of our study is to demonstrate the use of completely automatic objective measurements performed over multiple sentence-long utterances in the investigation of speaker identity perception. In the following sections, we present a description of the objective measurement set we use and give results of tests performed on the TIMIT speakers to determine the speaker discrimination merit of each objectively measurable descriptor. Next, their perceptual relevance is investigated using MDS on the subjective dissimilarity ratings between speaker pairs from a separate data set, which had been collected specifically for speaker recognizability testing.

2. OBJECTIVE MEASUREMENTS

The objective measurements used to investigate the problem of speaker identity perception by humans should naturally be related to the speech production process. Therefore, the physiological and prosodic attributes of the general discrete-time speech production model are useful as potential objectively measurable descriptors. Table 1 gives a listing of the objective measurements evaluated for this study. A frame-by-frame 10^{th} order linear predictive (LP) analysis, and the pitch and energy contours are the starting points for all the given objective measurements. The average vocal tract length is computed by fitting a uniform tube approximation to the formant structure of every voiced

speech frame [12, 13]. For this purpose, the required formant frequency estimates are obtained using the complex poles from the LP analysis. The glottal pulse prototype is obtained by filtering the voiced segments of the speech waveform with a reduced order inverse filter to remove the formant effects [3]. In addition to what we reported before, the obtained glottal pulse prototype is integrated to remove the lip radiation effect. Figure 1 shows example glottal pulse prototypes for a male and a female speaker from the TIMIT database. Opening and closing instants are determined as the points in time where 5% of the maximum attained level is crossed, and similarly, the top (open) portion is taken to be the region between the two points in time where 95% of the maximum level is reached during opening and closing phases. The spectral tilt is given as the average slope of the glottal pulse prototype spectrum magnitude.

Table 1. List of evaluated objective measurements.

Measurement	Description
POL{1...5}-mag POL{1...5}-ang	Magnitude and angle averages of complex poles from the 10^{th} order LP analysis of each voiced frame
VLEN	Vocal tract length estimate
GPP-t-open	Opening duration of the glottal pulse prototype (GPP)
GPP-s-open	Opening slope of GPP
GPP-peak	Peak level of GPP
GPP-t-peak	Time of GPP peak
GPP-t-top	Top (open) duration of GPP
GPP-s-close	Closing slope of GPP
GPP-t-close	Closing duration of GPP
GPP-tilt	Spectral tilt of GPP (dB/Octave)
PCH	Median pitch frequency
PCH-R	Pitch frequency range (90% of range centered at the median)
UV-SEGD	Average duration of *unvoiced* segments
VO-SEGD	Average duration of *voiced* segments
ENG-UV	Average energy of *unvoiced* segments
ENG-UV-R	Energy range of *unvoiced* segments (90% of range centered at the median)
ENG-VO	Average energy of *voiced* segments
ENG-VO-R	Energy range of *voiced* segments (90% of range centered at the median)

The speaker discrimination merit of the given objective measurements is evaluated through the maximum likelihood (ML) same/different classification of distances between speaker pairs for each descriptor [3]. From the TIMIT Continuous Speech Corpus, 86 male and 78 female speakers are selected so that within each gender group, the sentences for each speaker are unique, except for the two dialect calibration sentences common to all TIMIT speakers. Each gender group is divided into two subgroups for use as training and test sets. With this selection, sentence pairings are constructed to obtain a total of 2408 (86 speakers$\times 8 \times 7/2$) same-speaker pairs and 14448 (2 subgroups$\times 43 \times 42/2 \times 8$ sentences) different-speaker pairs for male speakers; and 2184 (78 speakers$\times 8 \times 7/2$) same-speaker pairs and 11856 (2 subgroups$\times 39 \times 38/2 \times 8$ sentences) different-speaker pairs for female speakers.

Following the construction of the same- and different-speaker pairs sets, the $3700Hz$ bandlimited speech waveforms for each sentence (average duration of $3sec$) were analyzed over $20msec$ Hamming windowed frames repeated every $10msec$, and distances between the utterances from speaker pairs were computed to generate the sets of same-speaker distances and different-speaker distances. For the two subgroups of speakers for each gender, the probability density functions (PDF) for the classes of same- and different-speaker distances were estimated separately for each objectively measured descriptor, so that the ML classification of distances from each subgroup could be performed using PDF parameters estimated on the other subgroup. The error rates obtained by the ML classification of distances is given in Table 2, and they represent the average of two train/test conditions. The distance between the complex poles from two sentences is a modified Mahalanobis distance such that

$$\begin{aligned} d^2(v_1, v_2) &= .5(v_1 - v_2)^t (C_1^{-1} + C_2^{-1})(v_1 - v_2) \\ &\quad + .5\mathrm{Tr}\{C_1^{-1} C_2 + C_1 C_2^{-1}\} - 10, \end{aligned}$$

where v_1 and v_2 are the 10-dimensional vectors composed of the averaged real and imaginary parts of the complex poles on the upper half plane, and C_1 and C_2 are the corresponding estimated covariance matrices. This is a measure of the distance between two clusters, and since the complex poles from the LP analysis of each voiced frame form clusters on the complex plane, this distance is a sensible choice, and indeed, it achieves the best error rate when compared with other distance measures considered between complex poles. The histogram and Gamma distribution PDF estimates for the complex pole cluster distances is shown in Figure 2. As can be observed, the Gamma distribution fits the distance data quite well. The distances for all the other measurements are given either as their absolute difference, or the absolute difference of their logarithms. For each descriptor, the distance achieving the minimum error rate is displayed. When compared with our previous results [3], the error rates seem rather high for the descriptor distances common to the two studies, but it should be noted that those results were obtained using roughly $20sec$ of speech for a measurement while this study involved $3sec$ long sentences. Yet, simply combining the likelihood scores of the distances marked

Figure 1. Example male and female speaker glottal pulse prototypes in time and frequency domains.

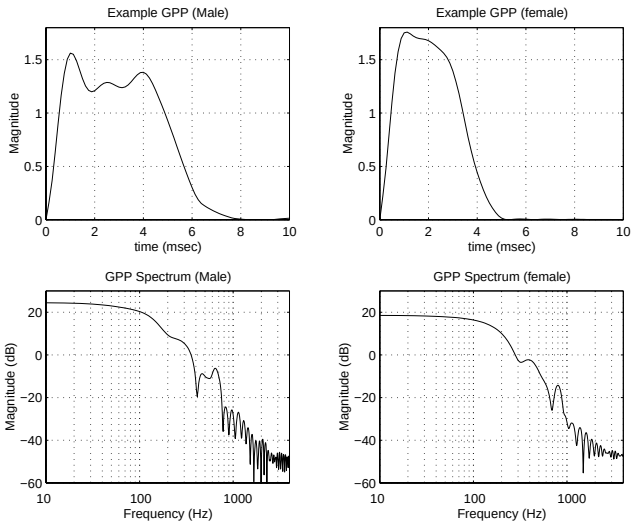
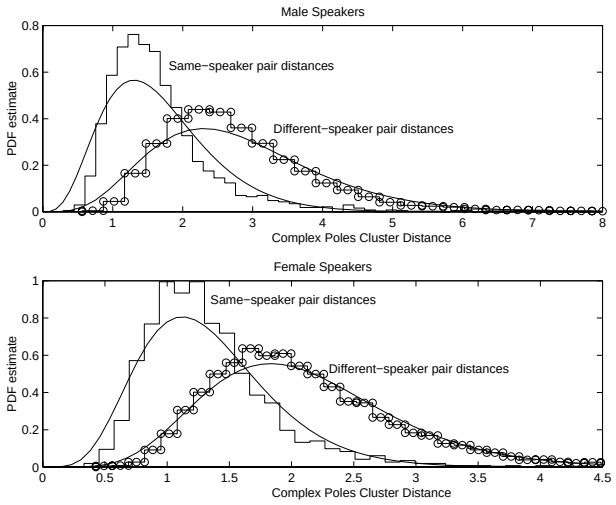


Figure 2. Histograms and Gamma distribution estimates for complex pole cluster distances between same and different speaker pairs from TIMIT.



with * achieves an error rate as low as $4.13 \pm .49\%$ for male speakers and $8.17 \pm .67$ for female speakers, and shows that these objectively measurable descriptors could still be potentially useful in speaker discrimination, using only $3sec$ of speech for each measurement.

Table 2. Maximum likelihood same/different speaker classification errors (with 95% confidence intervals) for TIMIT speaker pairs. The distance function used for each measurement is given in parentheses.

Objective Measure	Male (%)	Female (%)
*POLES (cluster)	23.61 ± .86	25.85 ± .96
*VLEN (log)	41.27 ± 1.05	35.45 ± 1.05
*GPP-peak (log)	45.15 ± 1.06	39.76 ± 1.06
GPP-s-close (log)	38.81 ± 1.06	36.06 ± 1.11
GPP-s-open (log)	47.81 ± .75	41.63 ± 1.12
*GPP-t-close (abs)	29.37 ± 1.02	35.54 ± 1.12
*GPP-t-open (abs)	24.85 ± 1.00	30.56 ± 1.07
GPP-t-peak (abs)	27.60 ± .99	32.67 ± 1.09
*GPP-t-top (abs)	27.20 ± 1.00	37.69 ± 1.12
*GPP-tilt (log)	40.15 ± 1.02	38.13 ± 1.08
(GPP all)	8.75 ± .73	18.51 ± .97
*PCH (log)	24.93 ± .89	24.63 ± .98
PCH-R (log)	40.85 ± 1.07	42.09 ± 1.13
*UV-SEGD (log)	46.71 ± 1.07	46.07 ± 1.14
*VO-SEGD (log)	50.12 ± 1.02	48.62 ± 1.12
*ENG-UV (log)	40.69 ± 1.03	43.65 ± 1.10
ENG-UV-R (log)	43.98 ± 1.03	44.85 ± 1.11
*ENG-VO (log)	32.11 ± .96	31.00 ± 1.01
ENG-VO-R (log)	34.51 ± .98	33.38 ± 1.04
*'s combined	4.13 ± .49	8.17 ± .67

3. SUBJECTIVE EVALUATION

Perceptual relevance of the described objective measurements were tested on a separate data set of speakers, which was collected for the purpose of speaker recognizability testing during the selection process for a new 2400 bps DoD standard coder [2]. The motivation to use this data set was the readily available subjective dissimilarity ratings and same/different judgments by 80 listeners — a much larger number of listeners compared with previous studies involving subjective dissimilarity judgments between utterance pairs.

The data set used for the subjective evaluation of objectively measured descriptors included ten male and ten female speakers, each with 36 unique sentences. Similarity judgments of 90 same-speaker pairs and 90 different-speaker pairs by each of the 80 listeners were available for clean, unprocessed speech, which was collected as part of the coder evaluation process (the sentence pair construction methods and other details can be found in the original paper by Schmidt-Nielsen and Brock [2].) The equal number of same and different pairs was expected to cause minimal bias for the response strategies of the listeners. Table 3 gives a brief summary of the response statistics for the listeners. (Each listener rated every speaker pair as same (0) or different (1) and then gave a judgment of similarity — (0) for very similar and (4) for very dissimilar.) The listener error rates show that listeners in general have more difficulty when judging female speaker pairs. The same combination of objective measurements (marked * in Table 2) achieved a ML distance classification error rate of $10.00 \pm 4.38\%$ for male speakers, and $17.22 \pm 5.51\%$ for female speakers of this subjective evaluation data set when PDF parameters were estimated using the distance data from both TIMIT subgroups for each gender.

Table 3. Statistics from the subjective similarity judgment tests.

	Male spkr. pairs		Female spkr. pairs	
Mean judgment	Same	Different	Same	Different
Same/Different	.0853	.9619	.2292	0.7285
Dissimilarity	.4493	2.9446	.9158	2.5711

Listener error	Male spkr. pairs	Female spkr. pairs
Lowest	1.1% (2/180)	6.1% (11/180)
Median	4.4% (8/180)	15.0% (27/180)
Highest	22.8% (41/180)	38.3% (69/180)

To construct a space of perceptional dimensions, Kruskal's multidimensional scaling was performed using the average subjective dissimilarity judgments. The resultant 3-dimensional configurations had stress levels of 5.222% for male and 6.671% for female speakers, which can be regarded as a "good" fit for the monotonic relationship between subjective similarity judgments and the corresponding distances in the solution space [11]. Table 4 displays those objectively measured descriptors which demonstrate a correlation at the .01 significance level with at least one of the orthogonal MDS dimensions for male or female speakers. The first dimension for male speakers seems to have a correlation with vocal tract features (average vocal tract length and the angle of the average fourth pole) and median pitch. The third dimension is correlated with the magnitude of the fifth complex pole averages which may be related to the bandwidth of the higher formants within the $3700Hz$ bandwidth of speech, while the second dimension does not display any significant correlation with any of the objective measurements. In the case of female speakers glottal and prosodic features (the spectral tilt computed from the

glottal pulse prototype as well as median pitch and average unvoiced segment duration) seem to have a very high correlation with the first dimension, while the second dimension displays a weaker correlation (at .05 level) with the magnitude of the fifth complex pole averages. For both males and females, there seems to emerge a dimension which has no correlation with events measurable with the evaluated objective descriptors, and although not equally significant for both cases, fifth complex pole average magnitudes seem to be correlated to a second perception dimension. The dimension of highest perceptual variance for both male and female speakers appear to be correlated with median pitch, and this dimension is also correlated with vocal tract features for male speakers and glottal and prosodic features for female speakers.

Table 4. Estimated correlation coefficients between the MDS solution dimensions and objective measurements (*Significant at .05 / $^\bullet$Significant at .01)

Objective Measure	Male Speakers			Female Speakers		
	D_1	D_2	D_3	D_1	D_2	D_3
GPP-tilt	-.51	-.34	.48	.93$^\bullet$	-.08	-.21
POL4-ang	-.84$^\bullet$	.37	-.16	-.21	.41	-.15
POL5-mag	-.14	.14	.85$^\bullet$	.51	.68*	-.33
VLEN	.77$^\bullet$	-.39	.26	.36	.14	.39
PCH	-.78$^\bullet$	-.52	-.14	.83$^\bullet$	-.28	.07
UV-SEGD	.45	.14	.01	-.77$^\bullet$	.26	.10

4. CONCLUSIONS

We have evaluated the speaker discrimination merit of a set of objectively measurable descriptors using single sentence utterances with an average duration of just 3*sec*. The results show that although the individual ML same/different speaker classification performances for most descriptors are not strong, their combinations can achieve very good detection rates when tested on the TIMIT speakers, and perform close to the median speakers when tested on the subjective evaluation set speakers, in spite of the small size of available data to perform measurements. However, being automatically measured objective descriptors, it is also possible to use them with larger amounts of data to achieve better discrimination potential and, furthermore, make the subjective perception studies with larger number of speakers and longer duration sentences more feasible by diminishing the need for hand measurements.

The 3-dimensional MDS solutions obtained from the subjective dissimilarity ratings show that median pitch is correlated with the main perceptual dimension for both male and female speakers, and the same dimensions show correlation with vocal tract features for male speakers, and glottal and prosodic features for female speakers. A second dimension seems to be correlated with the higher formant bandwidths for both genders, although for females, the correlation is weaker. None of the evaluated measures had a significant correlation with the third emerging perceptual dimension, and this should provide motivation for further investigation to develop new and improved objectively measurable descriptors.

However, the fact that not all the tested objective measurements showed significant correlations with the emerging perceptual dimensions for ten male and ten female speakers does not necessarily imply that they are potentially of limited use. Although a tested objective measure might indeed be perceptually irrelevant, there is also the possibility that the limited number of speakers precludes the emergence of other perceptual dimensions with which the measure might show correlation. A larger set of speakers, which will permit the emergence of more dimensions should be utilized to further test any potential objective measurements, especially those demonstrating strong merit in speaker discrimination.

REFERENCES

[1] M. H. L. Hecker and C. E. Williams, "On The Interrelation Among Speech Quality, Intelligibility, and Speaker Identifiability," *Proc. 5^e Congrés International D'Acoustique*, A15, Liége, September 1965.

[2] A. Schmidt-Nielsen and D. P. Brock, "Speaker Recognizability Testing For Voice Coders," *Proc. ICASSP'96*, Vol. II, pp. 1149–1152, Atlanta, GA, 1996.

[3] B. F. Necioglu, M. A. Clements and T. P. Barnwell III, "Objectively Measured Descriptors Applied to Speaker Characterization," *Proc. ICASSP'96*, Vol. I, pp. 483–486, Atlanta, GA, 1996.

[4] B. F. Necioglu, M. A. Clements and T. P. Barnwell III, "Reliability Assessment and Evaluation of Objectively Measured Descriptors for Speaker Characterization," *Proc. ICASSP'97*, Vol. II, pp. 955–958, Munich, Germany, 1997.

[5] W. D. Voiers, "Perceptual Bases of Speaker Identity," *The Journal of The Acoustical Society of America*, Vol. 36, No. 6, pp. 1065–1073, June 1964.

[6] W. D. Voiers, "Toward The Development of Practical Methods of Evaluating Speaker Recognizability," *Proc. ICASSP'79*, pp. 793–796, Washington, D.C., 1979.

[7] H. Matsumoto, S. Hiki, T. Sone and T. Nimura, "Multidimensional Representation of Personal Quality of Vowels and its Acoustical Correlates," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 5, pp. 428–436, October 1973.

[8] J. Kreiman, B. R. Gerrat, K. Precoda and G. S. Berke, "Individual Differences in Voice Quality Perception," *Journal of Speech and Hearing Research*, Vol. 35, pp. 512–520, June 1992.

[9] B. E. Walden, A. A. Montgomery, G. J. Gibeily, R. A. Prosek, and D. M. Schwartz, "Correlates of Psychological Dimensions in Talker Similarity," *Journal of Speech and Hearing Research*, Vol. 21, pp. 265–275, June 1978.

[10] S. Singh and T. Murry, "Multidimensional Classification of Normal Voice Qualities," *The Journal of The Acoustical Society of America*, Vol. 64, No. 1, pp. 81–87, July 1978.

[11] J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, No. 1, pp. 1–27, March, 1964.

[12] A. Paige and V. W. Zue, "Calculation of Vocal Tract Length," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-18, No. 3, pp. 268–270, September 1970.

[13] R. L. Kirlin, "*A Posteriori* Estimation of Vocal Tract Length," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, No. 6, pp. 571–574, December 1978.