RELIABILITY ASSESSMENT AND EVALUATION OF OBJECTIVELY MEASURED DESCRIPTORS FOR PERCEPTUAL SPEAKER CHARACTERIZATION

Burhan F. Necioğlu *Mark A. Clements* *Thomas P. Barnwell III*

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

ABSTRACT

With the more widespread use of lower bit rate speech coders, the evaluation of speaker recognizability becomes a major issue to be addressed as well as the evaluation of overall voice quality. Furthermore, subjective quality evaluation of speech coders may produce different results depending on the voice character of the speakers used in the evaluation process. It follows naturally that methods and procedures to characterize speakers perceptually must be devised. In this paper, we report on an enhanced set of objective descriptors of the speech waveform, assessing the reliability of their measurements as well as their merit in discriminating utterances from different speakers. Of the 45 measures presented, 35 have less than 10% RMS measurement error, and 25 of those have less than 5%.

1. INTRODUCTION

As the inevitable increase in the volume of traffic across the currently available communication channels dictates the need for more widespread use of lower bit rate speech coders, the evaluation of speaker recognizability becomes a major issue to be addressed as well as the overall quality evaluation of voice communication systems. This is attested by the fact that speaker recognizability was one of the requirements in the recent selection process for a new DoD 2400 bps voice coder standard [1]. Furthermore, subjective quality evaluation of speech coders may produce significant differences depending on the selection of speakers used in the evaluation process, suggesting that the voice characteristics of the test speakers act as an interference factor. This problem could be attended to, for instance, by a balanced selection of test speakers that yield similar "perceptual population profiles" among different test sets. It follows naturally that methods and procedures to characterize speakers perceptually must be devised.

In this paper, we report on our reliability assessment and merit evaluation of an *objective* descriptor set for speaker characterization, which is an enhanced version of the set we proposed previously [2]. The TIMIT Continuous Speech Corpus is used for the actual test and evaluations. The goal is to produce a set of objective measurements with a high potential for discriminating between different speakers while at the same time sustaining a high level of reliability, or repeatability. The next phase of this research effort will consist of a joint analysis of these objective measurements with a set of subjective speaker dissimilarity ratings, paving the way toward the eventual goal of predicting perceived speaker character through objective measurements.

2. OBJECTIVE MEASURES

The problem of speaker recognizability is complicated by the fact that perceptual characterization is a highly subjective phenomenon, with no established rules, and furthermore, with a potentially wide range of perceptual judgments. It is not well understood how humans process speech to extract identity information, or what kind of traits they utilize to characterize speakers perceptually. Due to the nature of the speech production process, the problem has a physiological component, which includes features such as glottal characteristics, vocal tract shape and length, and a prosodic component, which includes features such as vocal gestures, accent, speaking rate, pitch, modulation, and so forth. Therefore, the investigation of the contributing factors to voice character and identity would require extracting and measuring parameters related to both the physiological and prosodic features of the speech signal. These measurements can be subjective, with human listeners rating voices on a predetermined set of scales, as can be found in studies by Voiers [3], or can be objective, as we have previously reported, utilizing signal processing algorithms to extract information [2]. However, it is very likely that a robust speaker recognizability test will have to utilize both objective and subjective descriptors, since it might not be possible to subjectively assess some of the phenomena that can be measured objectively, and vice versa.

In Table 1 the set of objective measurements which are evaluated for this study are listed in three major groups of measurements related to *prosodic,vocal tract,* and *glottal* features. This set is an enhanced version of the set of objective measurements we reported previously in [2]. The *prosodic* measurements depend on the energy and pitch contours of a speaker's speech waveform, while *vocal tract* characteristics depend on a 10^{th} order Linear Predictive (LP) analysis. The *glottal* features are extracted utilizing both the averaged frame spectra and the averaged pseudo glottal waveform, which also depends on LP analysis, as described in [2]. It should be considered that, although many of the objective criteria considered are features that can be used for speaker verification and/or identification, they are not optimized for those purposes. Rather, we extract parameters intended to characterize speakers mechanically and

Table 1. List of evaluated objective measurements.

PROSODIC FEATURES	
LOG-EN-AVG	Log-energy average of all speech frames
LOG-EN-MAX	Maximum log-energy of all speech frames
EN-SDEV	Standard deviation of energy of all speech frames
EN-RANGE	Energy range of all speech frames
EN-AVG-MED	Difference between average and median energy of all frames
VO-LOG-EN-AVG	Log-energy average of voiced speech frames
VO-LOG-EN-MIN	Minimum log-energy of voiced speech frames
VO-LOG-EN-MAX	Maximum log-energy of voiced speech frames
VO-EN-SDEV	Standard deviation of energy of voiced speech frames
VO-EN-RANGE	Energy range of voiced speech frames
UV-LOG-EN-MED	Median log-energy of unvoiced speech frames
UV-LOG-EN-MAX	Maximum log-energy of unvoiced speech frames
UV-EN-SDEV	Standard deviation of energy of unvoiced speech frames
UV-EN-RANGE	Energy range of unvoiced speech frames
P-AVG	Average pitch period
P-SDEV	Standard deviation of pitch period
P-MED	Median pitch period
P-RANGE	Pitch period range
P-MIN	Minimum pitch period
PF-MIN	Minimum pitch frequency
SRATE	Speaking rate estimate (voiced-to-unvoiced-transitions/second)
UV-SEGD	Average duration of unvoiced speech segments.
VOCAL TRACT FEATURES	
POL{1, . . . , 5}-MAG	Magnitude averages of complex poles from LP analysis of each speech frame
POL{1, . . . , 5}-ANG	Angle averages of complex poles from LP analysis of each speech frame
VLEN	Vocal tract length estimate
VLEN-SDEV	Standard deviation of vocal tract length estimate
PGAIN	Average prediction gain
GLOTTAL FEATURES	
GPP-POW	Power of glottal pulse prototype (GPP)
GPP-RISE	Rise time to peak of GPP
GPP-SL-1	Rise slope of GPP
GPP-SL-2	Fall slope of GPP
GPP-F{1,2}	1^{st} and 2^{nd} major harmonic component frequencies of GPP
GPP-M{1,2}	Magnitudes of 1^{st} and 2^{nd} major harmonic components of GPP
GPP-TILT	(GPP-F1 − GPP-F2)/(GPP-M1 − GPP-M2)
STILT	Spectral tilt estimate from the averaged voiced segment spectra
STILT-MSE	Mean squared error of the spectral tilt estimate

perceptually. For example, the Glottal Pulse Prototype approximation (GPP), which is a finite duration signal template, is not evaluated by itself as a whole. Instead, several scalar features, considered to be representative of its various properties are computed for evaluation, both in time and frequency domain. Also added to the set of measurements are more detailed statistics extracted from the energy and pitch contour of a speaker's speech waveform, including separate energy statistics for both the voiced and unvoiced speech segments, all with the goal of capturing as much prosodic information as possible.

3. RELIABILITY ASSESSMENT AND EVALUATION

One of the major requirements of any measurement of speaker recognizability, whether subjective or objective in nature, is repeatability, which is a prominent factor in the formulation of descriptor sets [4]. The quantity we use as the reliability figure for a particular descriptor is the cross-correlation of two measurements of that descriptor on different sets of sentences spoken by a speaker.

With the assumption of independent and identically distributed observations for the additive Gaussian "noise" to a measurement, this reliability figure is simply $1 - e^2$

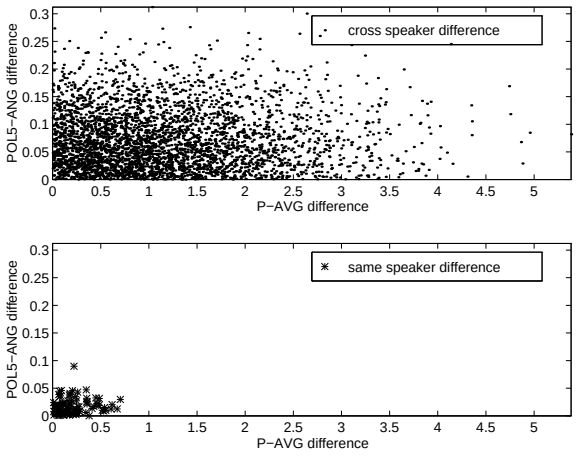
Table 2. Cross-correlations of objective measurements on two different speech waveforms of a speaker, RMS percentage error of measurements, and cluster quality measures (CQM) for measured differences between speakers.

Measurement	Cross-correlation	RMS error (%)	CQM
LOG-EN-AVG	0.9984	4.03	0.216
LOG-EN-MAX	0.9994	2.35	0.344
EN-SDEV	0.9777	15.11	0.333
EN-RANGE	0.9738	16.40	0.281
EN-AVG-MED	0.9575	21.07	0.221
VO-LOG-EN-AVG	0.9995	2.33	0.390
VO-LOG-EN-MIN	0.9990	3.14	0.354
VO-LOG-EN-MAX	0.9994	2.52	0.322
VO-EN-SDEV	0.9713	17.18	0.300
VO-EN-RANGE	0.9656	18.88	0.251
UV-LOG-EN-MED	0.9980	4.52	0.203
UV-LOG-EN-MAX	0.9993	2.73	0.180
UV-EN-SDEV	0.9743	16.23	0.164
UV-EN-RANGE	0.9777	15.10	0.149
P-AVG	0.9995	2.29	0.566
P-SDEV	0.9816	13.70	0.108
P-MED	0.9995	2.32	0.562
P-RANGE	0.9729	16.68	0.138
P-MIN	0.9955	6.73	0.310
PF-MIN	0.9980	4.49	0.404
SRATE	0.9932	8.29	0.021
UV-SEGD	0.9838	12.83	0.023
POL1-MAG	1.0000	0.53	0.154
POL1-ANG	0.9988	3.41	0.072
POL2-MAG	0.9992	2.83	0.304
POL2-ANG	0.9986	3.70	0.241
POL3-MAG	0.9992	2.89	0.141
POL3-ANG	0.9993	2.59	0.282
POL4-MAG	0.9992	2.74	0.226
POL4-ANG	0.9998	1.54	0.342
POL5-MAG	0.9998	1.51	0.134
POL5-ANG	1.0000	0.63	0.469
VLEN	0.9994	2.52	0.161
VLEN-SDEV	0.9870	11.48	0.106
PGAIN	0.9740	16.34	0.216
GPP-POW	0.9981	4.34	0.240
GPP-RISE	0.9956	6.68	0.029
GPP-SL-1	0.9939	7.83	0.078
GPP-SL-2	0.8787	37.15	0.155
GPP-F1	0.9981	4.33	0.272
GPP-F2	0.9978	4.73	0.328
GPP-M1	0.9978	4.74	0.301
GPP-M2	0.9957	6.56	0.274
GPP-TILT	0.9134	30.79	0.318
STILT	0.9923	8.78	0.149
STILT-MSE	0.9937	7.94	0.106

where $e = \sqrt{\sigma^2/(m^2+\sigma^2)}$ is the normalized root mean squared (RMS) measurement error, with σ^2 denoting the noise power and m the measured quantity.

We evaluate the merit of our objective measurements in a speaker-identity discrimination context, using an invariant criterion of cluster scattering [5] for the classes of *difference* data as in [2]. The objective measure differences between utterances of speakers are generated, forming two classes: The *same-speaker differences* and the *cross-speaker differences*. An objective measure with a better potential for the discrimination of *different* speakers and the detection of *same* speakers should have a better separation between the *same-* and *cross-speaker* difference clusters with a low variance, or scatter, within each cluster. The Cluster Quality Measure (CQM), which is given by $tr\{C_W^{-1}C_B\}$, where C_W and C_B denote the within-class and between-class covariance matrices respectively, is therefore used as a figure of merit.

Figure 1. Cluster plot of measured P-AVG differences versus measured POL5-ANG differences for *same*- and *cross-speaker* classes.



4. EVALUATION EXPERIMENTS

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus was used for the actual test and evaluations of the objective measurements presented in this study. In our previous study, we had used a data set of 80 male speakers all uttering the same 16 sentences [2]. This time we elected to use a subset of 86 male speakers from the TIMIT corpus, constructed such that the transcription of each speaker's sentences were unique to that speaker, except for the two dialect sentences common to all speakers. This allowed us to perform and compare our measurements on a rich and diverse pool of acoustic data.

For each of the 86 speakers, and each proposed objective descriptor, separate measurements were performed over two 4-sentence long utterances which had an average duration of approximately 12 seconds. All frame based computations were performed every 10 *msec*, on 20 *msec* long Hamming windowed overlapping segments of the speech waveform which was downsampled to 8 *kHz* from the original 16 *kHz*. A total of 86 *same-speaker* and 3655 *cross-speaker* comparisons were performed for each of the given objective descriptor measurements.

Table 2 gives a list of all 45 objective measurements with their respective computed reliability and merit figures. Thirtytwo of the measurements have an RMS measurement error lower than 10% and they include measurements from all three major groups of prosodic, vocal tract, and glottal features, and 25 of those display errors lower than 5%. We hypothesize that the descriptors with larger error figures might be suffering from insufficient data, rather than being "unreliable" descriptors. Of course, there is the possibility that the quantity being measured has a relatively large variance itself, and should not be used as a measure of perceived identity or character. However, testing these hypotheses requires a substantial amount of data *per* speaker, which is not available to us at this time.

In Table 3, we give the computed CQM's for combined objective measurement differences. Starting with the objec-

Figure 2. Cluster plot of measured POL4-ANG differences versus measured POL5-ANG differences for *same*- and *cross-speaker* classes.

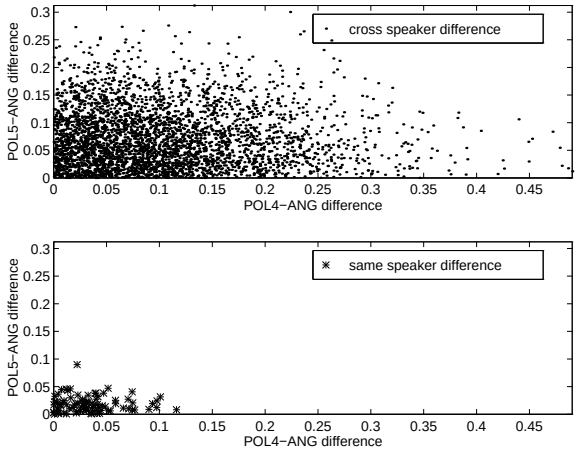


Figure 3. Cluster plot of measured P-AVG differences versus measured POL5-ANG differences versus measured POL4-ANG differences for *same*- and *cross-speaker* classes.

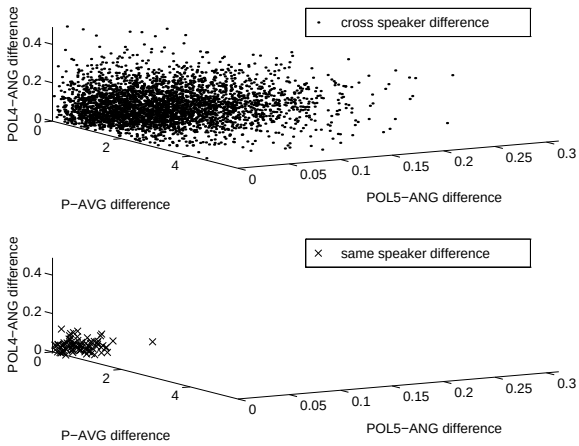


Table 3. CQM figures for combined measurement differences. At each line the measurement that enhances the combined CQM most has been added to the list of preceding descriptors.

Combined Measurements	Combined CQM	Increase (%)
P-AVG	0.566	-
+ POL5-ANG	1.022	80.71
+ POL4-ANG	1.346	31.60
+ VO-LOG-EN-AVG	1.630	21.10
+ GPP-TILT	1.925	18.14
+ POL2-ANG	2.148	11.60
+ GPP-SL-2	2.295	6.82
+ GPP-POW	2.438	6.23
+ POL3-MAG	2.571	5.47
+ POL1-MAG	2.672	3.90
+ VLEN-SDEV	2.781	4.12
+ POL2-MAG	2.860	2.81
+ POL5-MAG	2.920	2.11
+ P-SDEV	2.963	1.46
+ UV-LOG-EN-MED	2.993	1.03
+ VLEN	3.023	0.99
+ GPP-SL-1	3.056	1.12
+ PGAIN	3.087	1.02
+ SRATE	3.118	1.00
+ LOG-EN-AVG	3.134	0.51
+ EN-SDEV	3.148	0.45
+ VO-EN-RANGE	3.200	1.63
+ VO-EN-SDEV	3.222	0.69
+ P-RANGE	3.238	0.50
+ POL4-MAG	3.249	0.36
+ STILT-MSE	3.259	0.28
+ EN-RANGE	3.267	0.24
+ VO-LOG-EN-MIN	3.273	0.18
+ POL1-ANG	3.278	0.17
+ GPPm2	3.284	0.17
+ LOG-EN-MAX	3.287	0.10
+ VO-LOG-EN-MAX	3.311	0.74
+ P-MED	3.314	0.10
+ GPP-RISE	3.317	0.08
+ PF-MIN	3.320	0.09
+ GPP-F1	3.323	0.09
+ UV-EN-SDEV	3.325	0.06
+ EN-AVG-MED	3.327	0.06
+ POL3-ANG	3.329	0.05
+ STILT	3.330	0.05
+ P-MIN	3.332	0.04
+ GPP-M1	3.333	0.03
+ UVSEGD	3.334	0.03
+ UV-LOG-EN-MAX	3.334	0.02
+ UV-EN-RANGE	3.339	0.15

tive measurement with the best individual CQM, the next measurement which causes the highest increase in the combined CQM is added to the combination list at each step. With this suboptimal search, the objective measures are ordered, together with their respective CQM's for the combined differences. Combining all 45 measures achieves a max CQM of 3.34, and 90% of this maximum is achieved by the first 15 measures, of which 8 are related to vocal tract features, 4 to prosodic features and 3 to glottal features.

5. CONCLUSIONS

In this paper, we have presented an enhanced set of objective measurements potentially useful for perceptual characterization of human speech. The measurements are related to three major groups, which consist of prosodic, vocal tract and glottal features.

Of the 45 objective measures presented, 32 have less than 10% RMS measurement error, and only three of the measurements have an RMS measurement error between 20%-40%. The CQM figures for combined objective measure differences show that certain measurements from all three major groups contribute to maximize the cluster quality when incremental combination and CQM computation of differences is performed.

The 2- and 3-dimensional cluster plots of the top three objective measurements from Table 3 in Figures 1, 2 and 3 offer a visual interpretation of the computed CQM quantities. The concentration of the class of *same*-speaker differences near the origin is clearly visible in these plots. However, although the mean (center) point of the class of *same*-speaker differences appear quite separated from the mean point of the class of *cross-speaker* differences, the scatter of the latter class is also clearly present, which almost engulfs the region of the former. Actually, this may not be a weak point, but an indicator of the fact that different speakers *may* indeed sound like each other. A joint analysis of these objective measurements with accompanying subjective dissimilarity data must be performed in order to determine the *real* merit of an objective difference measure as an indicator of *perceived* identity. This aspect of the problem will be investigated in the next phase of our research, toward the eventual goal of predicting if not all but most parts of perceived speaker character through objective measurements.

REFERENCES

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

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

[3] W. D. Voiers, "Toward The Development of Practical Methods of Evaluating Speaker Recognizability," *Proc. ICASSP'78*, pp. 793–796, Washington, D.C., 1978.

[4] W. D. Voiers, Personal Communication, 1996.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, 1973).