Effects of Audio and ASR Quality on Cepstral and High-level Speaker Verification Systems

Andreas Stolcke

Conversational Systems Lab
Microsoft
Mountain View, CA, U.S.A.
anstolck@microsoft.com

Martin Graciarena *Luciana Ferrer*

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA, U.S.A.
martin,lferrer@speech.sri.com

Abstract

Speech data for NIST speaker recognition evaluations has traditionally been distributed in compressed, telephone quality form, even for microphone data that was originally recorded at higher quality. We evaluate the effect that improved audio quality has for speaker verification performance, using a recently released full-bandwidth version of microphone data from the SRE2010 evaluation. Remarkably, we find substantially improved results even though the underlying speaker recognition models remain based on a telephone-band feature front end. For a cepstral GMM system we show improvements purely from the elimination of lossy (μlaw) coding and more effective noise reduction filtering at the full bandwidth. We also find that higher-level speaker recognition systems can benefit from better ASR quality enabled by the improved audio quality. Specifically, we show that a speech recognizer trained on full-bandwidth, distant-microphone meeting speech data yields reduced speaker verification error for speaker models based on MLLR features and word-N-gram features.

1. Introduction

Much of the progress in speaker recognition has been driven by annual or biannual open technology evaluations administered by the U.S. National Institute of Standards and Technology (NIST) [1, 2]. Historically, the data for these Speaker Recognition Evaluations (SREs) has been collected as conversations over the telephone. Starting in 2005, a relatively small number of trials involved telephone calls where one side was also recorded over microphones [3], and the two most recent evaluations (2008 and 2010) made use of substantial amounts of data from a new genre: one-on-one interviews in an office setting and recorded by a variety of microphones and at different distances from the speaker [4]. However, even though these microphone recordings were originally digitized at higher sampling rates, they continued to be delivered to system developers in telephone quality: downsampled to 8 kHz and encoded in lossy 8-bit μlaw. Presumably, this was done to lessen the burden on developers to deal with bandwidth and coding mismatch, since much of the background training data continued to be drawn from telephone sources. (A secondary consideration, at least when disk space was at a higher premium, might have been that the 8-bit coding cut storage requirements for the data in half.)

This historical perspective, and the greater emphasis in recent evaluations on nontelephone data, raises the question: how much benefit could be derived from having SRE data available at its full bandwidth, without lossy encoding? Inspection of SRE data showed that μlaw coding is especially problematic for low-energy speech, such as might be found with distant microphone recordings, since μlaw effectively only uses a few bits to encode small amplitudes.

After the 2008 evaluation, for unrelated reasons, we obtained a wide-band version of the interview subset of the evaluation data, and ran preliminary experiments (reported at the SRE2010 workshop in Brno) showing that our systems could benefit substantially from use of the higher audio quality, and argued for a change in SRE practice so that, in the future, participants would receive microphone data in 16-bit PCM encoding, sampled at least at 16 kHz. Recently, NIST has re-released the SRE2010 microphone data in this form, allowing us to validate the earlier results using the most recent evaluation data and the full SRE2010 extended trial set.

In this paper, we summarize the results on SRE2008 data, and report the effect of audio encoding and attendant changes in automatic speech recognition (ASR) quality on a subset of our evaluation systems on the full-band SRE2010 data. The results again show substantial benefits for the performance of speaker verification systems, both standard cepstral systems and those systems based on higher-level information derived from ASR.

Research was done in part while first author was with SRI International.

Table 1: SRE2008-based development set statistics

Train-test condition	Target trials	Impostor trials
short-short.int-int.mic-mic	33,743	1,108,882
short-long.int-int.mic-mic	10,234	336,437
long-short.int-int.mic-mic	32,248	1,054,592
long-long.int-int.mic-mic	9,774	319,956

2. Method

2.1. Data and error metrics

We used two datasets for our experiments. The first, older one is a development set assembled from SRE2008 data. A set of 82 interview speakers (48 females and 34 males) was held out from the SRE2008 trial definition to be used for training the systems. These were all speakers from the interview conditions, some from the original set, others from the "follow-on" set distributed after SRE2008. A development test set was then created using the remaining SRE2008 data. For each original condition from SRE2008 an extended set was created by pairing every available model against every available test sample (except when the model and the test sample used data from the same original recording session). No additional models were created and only samples originally used for testing were used for testing in the extended set. The follow-on test data was added to all the interview conditions.

We had access to full-band versions only for the SRE2008 interview recordings, not the phonecall-over-microphone and (by definition) the phonecall-over-telephone recordings. Consequently, for the SRE2008 dataset, we report only on the interview-interview test condition, aggregating over two sample lengths: "short" (3 minutes) and "long" (8 minutes). Long interview waveforms from SRE2008 were truncated to 8 minutes to match the long interview samples in SRE2010. The resulting number of trials is shown in Table 1. This is the dataset we reported on at the SRE2010 workshop in Brno.

The second dataset was the re-released interview and phonecall-over-microphone data from SRE2010, evaluating on the "extended" trial set. The evaluation conditions were defined based on phonecall versus interview genre, vocal effort, and according to whether the same microphone was used in training and test. Table 2 summarizes the evaluation set and conditions.

We report results for three metrics: the traditional equal error rate (EER), which constrains false alarm and miss error rates to be the same, the old (pre-2010) detection cost function (oDCF), which weighs false alarm errors as ten times as costly as miss errors, and the new (2010) detection cost function (nDCF), which weighs false alarm errors as 1000 times more costly than miss errors. Old and new DCF values are scaled to make chance error rate equal to 1.

2.2. Waveform coding versions

The baseline waveform coding condition was that used by NIST. Interview and microphone data was downsampled (by NIST) to 8 kHz and 8-bit μlaw encoded, a lossy compressive coding commonly employed for telephone data. We call this the "8k-ulaw" version of the data.

For SRE2008 data we prepared two additional conditions. The 16 kHz Mixer-5 waveforms were cut to the same segments as used in SRE2008, and converted from FLAC to 16-bit PCM encoding. This yielded the "16k" version of the data. In addition, we downsampled this version to 8 kHz using the *sox* "polyphase" algorithm. Unlike NIST, we did not apply compressive coding and left the data in 16-bit PCM format. We call this the "8k" version of the data.

For SRE2010 date we compared only the "8k-ulaw" and the "16k" version of the data, both obtained directly from NIST.

2.3. Waveform preprocessing

All interview and microphone waveforms were noise-filtered with the Qualcomm-ICSI-OGI Aurora system implementation of a Wiener filter [5]. The goal of this step is to reduce ambient stationary noise picked up by the distant microphones. This implementation finds nonspeech regions using a neural-net classifier to estimate the noise spectrum. Appropriately trained versions of the filter are applicable to 8 kHz and 16 kHz waveforms. Following Wiener filtering, nonspeech regions are eliminated using a two-class HMM decoder, trained on telephone speech. We also used ASR output made available by NIST to eliminate regions with interviewer speech, to minimize crosstalk into the interviewee [6]. To avoid confounding our analyses with the effect of different waveform segmentations, we used the exact same segmentation points (namely, those used in the 8k-ulaw-based system for SRE2010) in all waveform versions. It is likely that the speech/nonspeech detection step would have benefited from the expanded bandwidth, but we did not want to introduce this additional variable into our comparisons.

2.4. Baseline verification system

As a baseline speaker verification system we chose the cepstral GMM-JFA subsystem in SRI's SRE2010 evaluation submission [6]. The cepstral GMM (Gaussian mixture model) system uses a 300-3300 Hz bandwidth front end consisting of 24 Mel filters to compute 20 cepstral coefficients with cepstral mean subtraction, and their delta and double delta coefficients, producing a 60-dimensional feature vector. The feature vectors are modeled by a 1024-component gender-independent GMM.

Table 2: SRE2010 evaluation set statistics, numbered according to NIST conditions. int = interview, mic = phonecall-over-microphone, nve = normal vocal effort, lve = low voc. eff., hve = high voc. eff. Conditions 1 and 2 were defined according to whether the microphone types in training and test where the same or different.

Train-test condition	Target trials	Impostor trials
01.int-int.same-mic	4,304	795,995
02.int-int.diff-mic	15,084	2,789,534
04.int-nve.mic-mic	3,637	756,775
07.nve-hve.mic-mic	359	82,551
09.nve-lve.mic-mic	290	70,500

The background GMM is trained using data from the 2004 and 2005 SRE and 2008 held-out interview data. Joint factor analysis (JFA) [7] is performed on mean supervector with speaker, channel and diagonal factors. Speaker factors are trained with 2004 and 2005 SRE data with additional data from the Switchboard-II corpus. Channel factors are obtained separately for telephone (phonecall and microphone) data and interview data. The two factors are combined to form single-channel factor matrix. The diagonal term is trained on the same data as the speaker factors. Scores are generated using asymmetrical scoring of subspace-adapted mean supervectors. The resulting scores are normalized using gender-dependent ZTnorm.

It is important to note that the Mel-cepstrum front end of the GMM-JFA system was configured for telephone data for all our experiments. This means that after all waveform processing, the front end effectively downsamples all 16 kHz data to 8 kHz. Otherwise, we would have had to deal with bandwidth mismatch in speaker modeling, a challenging research topic that was outside the scope of the present study. As a further expedient, we did not retrain the JFA (eigenspeakers and eigenchannels) in the model, simply retaining them from the 8k-ulaw version of the system.

2.5. Speech recognition systems

Beyond the cepstral baseline system, we also wanted to investigate waveform coding effects on higher-level speaker verification systems, specifically those based on ASR. We used two versions of SRI's conversational speech recognition system. The baseline version was the ASR system used in SRE2008 and SRE2010 and described in [6]. This system uses acoustic models trained exclusively on telephone speech, and runs in two recognition passes, for purposes of unsupervised adaptation. Microphone and telephone data use the same speech recognizer, although microphone data is first Wiener-filtered as described above. This system yields a word error rate (WER) of 28.8% on a hand-transcribed sample of SRE06 microphone data.

On 16k waveforms, a second ASR system was also tested. It is similar to the baseline system in structure and modeling algorithms employed, but uses a combination of 8 kHz and 16 kHz acoustic models, trained on both near-field and distant-microphone meeting recordings, with telephone and broadcast news data used as background training, respectively. The system represents the first two decoding passes of the SRI-ICSI meeting recognizer fielded in the NIST 2007 Rich Transcription (RT) evaluation [8].

We do not have transcribed SRE wide-band data on which to test the WER of the wide-band recognition system. When measured on meeting data (the single distant-microphone RT-07 conference meeting set), the two-pass telephone recognizer achieves 50.2% WER, the first pass of the meeting recognizer 42.3%, and the second pass of the meeting system 36.1%. On the SRE2010 interview data, we found that the meeting system recognized about 21% more word tokens than the telephone ASR system. This is a good indication of better ASR quality since our recognizer tends to delete, rather than misrecognize, words in mismatched conditions.

2.6. ASR-based verification systems

The second speaker modeling approach tested was a system based on maximum likelihood linear regression (MLLR) adaptation transforms modeled by support vector machines (SVMs) [9]. This system used a perceptual linear prediction (PLP) front end, with SVM features consisting of MLLR transforms specific to 8 phone classes and two gender-specific reference models (16 transforms in total), and nuisance attribute projection [10] for session variability compensation. This is the same MLLR subsystem as used in our evaluation system [6], but with ZT score normalization left out, to expedite experimentation and because it contributed little to performance.

As for the baseline system, we avoid issues of bandwidth mismatch by always employing 8 kHz acoustic models for MLLR feature estimation. However, the MLLR estimation can benefit from more accurate ASR hypotheses resulting from a wide-band ASR system. In fact, we tested three sets of ASR hypotheses: first-pass output from telephone ASR, first-pass output from meeting ASR, and second-pass output from meeting ASR. We did not find benefits from using second-pass telephone ASR for MLLR estimation. Note that the second pass of the meeting ASR system uses a 16 kHz front end, and gives about 15% relative WER reduction on meeting speech (RT 2007 single distant microphone condition) compared to the first pass, which is based on 8 kHz acoustic models.

A third verification system uses word N-gram frequency features, modeling the speakers' idiolects as proposed by [11]. It employs rank-normalized relative N-gram frequencies as features for SVM speaker models (for details see [6]). These speaker models are entirely based on ASR hypotheses, so we evaluate the system based on the different ASR systems described above. We always use the final (second pass) word hypotheses of the respective recognition system.

3. Experiments and Results

3.1. Cepstral system performance

We first evaluate the cepstral system on various waveform versions of the SRE2008 data. Table 3 shows the results and relative improvements over the baseline waveform condition (8k-ulaw).

The results show an overall error reduction on the order of 12% to 14% relative, depending on the metric used. The bulk of the gain comes from the dropping of μlaw coding, while a smaller share (2% relative) is the result of the higher sampling rate. It is worth remembering that the cepstral front end always operates at the lower, 8 kHz sampling rate. However, the noise filtering in preprocessing can take advantage of the expanded bandwidth, presumably by doing a better job at noise spectrum estimation.

Next, we compare 8k-ulaw and 16k waveforms only on the SRE2010 extended trial set, for all conditions involving microphones in training and test. Results are given in Table 5. The error reductions are smaller than on the development set, but are consistently positive. On conditions involving normal vocal effort, nDCF is reduced by about 7% and oDCF by about 7% to 10%. The conditions involving high and low vocal effort show some very large EER reductions, but the number of trials is very small. Still, it is suggestive that low-vocal-effort EER is reduced by 50%, given that μlaw coding is especially lossy at low amplitudes.

3.2. ASR-based system performance

Table 6 shows results for the various SRE2010 conditions for the MLLR-SVM system on SRE2010 data. For 16k waveforms, the meeting ASR system is employed, using either first- or second-pass hypotheses. Error reductions are generally around 15% relative for nDCF, 23% for oDCF, and around 20% for EER (with some outliers in conditions with low trial counts). Unlike that previously found for the telephone ASR system, the MLLR feature performance improves consistently when output from the second recognition pass is used. As for the cepstral system, we see an especially large EER reduction for the low-vocal effort condition.

Table 4 shows results for the word N-gram SVM system, comparing the 8k-ulaw baseline to the meeting-ASR based system run on 16k waveforms (both systems use second-pass ASR hypotheses). Note that only oDCF and EER results are given, as the word N-gram system gives nDCF values close to chance (1.0) for all conditions. The improvements from 16k waveform processing are much smaller than seen earlier, typically just a few percent relative. Still, the fact that the improvements are consistently positive shows that high quality audio and better ASR helps this system, too.

4. Conclusions and Further Work

We have shown that speaker verification on NIST microphone data (phonecall or the new interview genre) can benefit greatly from use of wide-band, lossless audio encoding, contrary to historical practice for NIST SRE data. This is true even without changing the bandwidth of cepstral feature extraction, and can be credited simply to the elimination of lossy coding and better noise filtering at the 16 kHz sampling rate. In addition, ASR-based speaker models, such as those based on MLLR and word N-gram features, get an additional boost from better ASR. While there is no SRE-like ASR training data, we found that a recognizer trained on distant microphone meeting recordings gives much improved results compared to a telephone speech recognizer.

We expect the availability of high-quality audio for future SREs to spur new advances in the field. Note that our experiments were suboptimal in that much of the older microphone data (from SRE05, SRE06, and the interview data released prior to SRE08) is still not available in wide-band form and could therefore yield further gains when models are properly retrained. The experiments reported here suggest various lines of future work. We have not yet evaluated the effect of better audio coding on other key elements of our SRE system, such as speech activity detection, cepstral GMM-JFA systems based on PLP, and prosodic speaker models [6]. It will be interesting to see what the overall improvement is after updating and combining all these systems. A more challenging, and ultimately interesting, question is what can be done to model the expanded audio bands in the cepstral feature space, while dealing with the issue of bandwidth mismatch between telephone and microphone data.

5. Acknowledgments

We thank Sachin Kajarekar and Nicolas Scheffer for developing the GMM-JFA baseline system, and for many helpful discussions.

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

Table 3: Cepstral GMM-JFA system performance on SRE2008 data, pooled over all interview-interview conditions. %chg refers to relative error reduction compared to the 8k-ulaw baseline.

Waveform	nDCF	%chg	oDCF	%chg	EER	%chg
8k-ulaw	.3300	-	.0506	-	.8140	-
8k	.3030	8.18	.0443	12.45	.7105	12.71
16k	.2920	11.52	.0434	14.23	.6977	14.29

Table 4: Word N-gram system performance on SRE2010 data, by evaluation condition

Condition	oDCF			EER		
	8k-ulaw	16k	%chg	8k-ulaw	16k	%chg
01.int-int.same-mic	.929	.920	0.94	29.2	28.3	2.94
02.int-int.diff-mic	.946	.929	1.79	31.7	29.4	7.17
04.int-nve.mic-mic	.942	.931	1.10	29.6	28.8	2.69
07.nve-hve.mic-mic	.868	.841	3.13	24.5	24.5	0.00
09.nve-lve.mic-mic	.887	.819	7.65	23.8	23.1	2.90

Table 5: Cepstral GMM-JFA system performance on SRE2010 data, by evaluation condition

Condition	nDCF			oDCF			EER		
	8k-ulaw	16k	%chg	8k-ulaw	16k	%chg	8k-ulaw	16k	%chg
01.int-int.same-mic	.4440	.4110	7.43	.0877	.0820	6.50	1.9284	1.8587	3.61
02.int-int.diff-mic	.5200	.4810	7.50	.1313	.1183	9.90	3.0894	2.7314	11.59
04.int-nve.mic-mic	.3900	.3610	7.44	.1074	.0970	9.68	2.2271	2.1721	2.47
07.nve-hve.mic-mic	.9020	.8700	3.55	.2381	.2045	14.11	4.7354	3.6212	23.53
09.nve-lve.mic-mic	.2980	.2010	32.55	.0575	.0502	12.70	1.3793	0.6897	50.00

Table 6: MLLR-SVM system performance on SRE2010 data, by evaluation condition. 16k(1) refers to the use of meeting ASR first-pass output, 16k(2) to second-pass output. %chg refers to 16k(2) error reduction over the 8k-ulaw baseline.

Condition	nDCF				oDCF				EER			
	8k-ulaw	16k(1)	16k(2)	%chg	8k-ulaw	16k(1)	16k(2)	%chg	8k-ulaw	16k(1)	16k(2)	%chg
01.int-int.same-mic	.5180	.4650	.4310	16.80	.2183	.1895	.1702	22.03	6.1338	5.3439	4.9954	18.56
02.int-int.diff-mic	.6500	.5620	.5370	17.38	.3259	.2687	.2494	23.47	9.6526	7.8427	7.3389	23.97
04.int-nve.mic-mic	.4940	.4620	.4260	13.77	.2181	.1840	.1677	23.11	6.2139	5.3066	4.8117	22.57
07.nve-hve.mic-mic	.8330	.8440	.7950	4.56	.3367	.3180	.2968	11.85	7.7994	8.0780	7.2423	7.14
09.nve-lve.mic-mic	.2660	.2600	.2970	-11.65	.1024	.0750	.0755	26.27	3.4483	3.1034	2.4138	30.00

6. References

[1] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective", *Speech Communication*, vol. 31, pp. 225–254, June 2000.

[2] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluation chronicles—part 2", *in Proceedings IEEE Odyssey-06 Speaker and Language Recognition Workshop*, pp. 1–6, San Juan, Puerto Rico, June 2006.

[3] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer corpora—2004, 2005, 2006", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1951–1959, Sep. 2007.

[4] A. F. Martin and C. S. Greenberg, "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels", *in Proc. Interspeech*, pp. 2579–2582, Brighton, U.K., Sep. 2009.

[5] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, vol. 1, pp. 4–7, Denver, Sep. 2002.

[6] N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system", *in Proc. ICASSP*, pp. 5292–5295, Prague, May 2011.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification", *in Proc. ICASSP*, vol. 1, pp. 113–116, Toulouse, May 2006.

[8] A. Stolcke, K. Boakye, Özgür Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system", in R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, vol. 4625 of *Lecture Notes in Computer Science*, pp. 450–463, Berlin, 2008. Springer.

[9] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987–1998, Sep. 2007.

[10] A. Solomonoff, C. Quillen, and I. Boardman, "Channel compensation for SVM speaker recognition", *in Proceedings Odyssey-04 Speaker and Language Recognition Workshop*, pp. 57–62, Toledo, Spain, May 2004.

[11] G. Doddington, "Speaker recognition based on idiolectal differences between speakers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, pp. 2521–2524, Aalborg, Denmark, Sep. 2001.