New Release of Mixer-6: Improved Validity for Phonetic Study of Speaker Variation and Identification

Eleanor Chodroff, Matthew Maciejewski, Jan Trmal, Sanjeev Khudanpur, John Godfrey
Johns Hopkins University
3400 N. Charles St., Baltimore, MD 21218
echodro1@jhu.edu, mmaciej2@jhu.edu, jtrmal@gmail.com, khudanpur@jhu.edu, godfrey.jack@gmail.com

Abstract

The Mixer series of speech corpora were collected over several years, principally to support annual NIST evaluations of speaker recognition (SR) technologies. These evaluations focused on conversational speech over a variety of channels and recording conditions. One of the series, Mixer-6, added a new condition, read speech, to support basic scientific research on speaker characteristics, as well as technology evaluation. With read speech it is possible to make relatively precise measurements of phonetic events and features, which can be correlated with the performance of speaker recognition algorithms, or directly used in phonetic analysis of speaker variability. The read speech, as originally recorded, was adequate for large-scale evaluations (e.g., fixed-text speaker ID algorithms) but only marginally suitable for acoustic-phonetic studies. Numerous errors due largely to speaker behavior remained in the corpus, with no record of their locations or rate of occurrence. We undertook the effort to correct this situation with automatic methods supplemented by human listening and annotation. The present paper describes the tools and methods, resulting corrections, and some examples of the kinds of research studies enabled by these enhancements.

Keywords: corpus, speech, transcription

1. Introduction

Speech corpora for research on speaker recognition (SR) are particularly challenging to design and collect. Best practices in this field call for accurate demographics, multiple sessions per speaker at certain intervals, control of speech style, signal quality, channels, transducers, and many other factors, all with accurate documentation.

The NIST Speaker Recognition evaluations (SRE) have provided de facto standards for such corpora over the last two decades, and the Mixer series of corpora were collected to meet SRE requirements over a period of several years (Cieri et al., 2004). They exercised a variety of conditions relevant to their intended applications: multiple languages, channels, environments, noises, speaking styles and many others. Mixer-6 followed the same protocols as the other Mixers in the collection of phone calls, but added two additional recording conditions of an interview and read speech portion (for a detailed description of the design and methods, see Brandschain et al., 2013).

For forensic, phonetic, and engineering research, this corpus can be very valuable, especially with regards to speaker identification and variability. In comparison to other spoken corpora, some of the distinguishing features of the Mixer-6 read speech include the sheer number of unique speakers and the amount of data collected per speaker, multiple recording sessions for most speakers, and the controlled sentence content and order. This design was primarily intended to advance research in forensic applications, i.e., SR decisions made by human experts for evidentiary use, with or without the aid of computer algorithms. The goal of this paper is to present an improved, audited version of the Mixer-6 read speech and to highlight its potential not only for forensic, but also for other phonetic and SR research.

2. The Original Release

Given the constraints of collecting three types of speech data per session from participants at a fixed total cost, LDC allocated a 15 minute interval for the read portion and had the speakers read as many sentences as possible. The sentences were prompted singly on a screen from a set list in a fixed order for all sessions. Participants managed to read an average of about 225 sentences per session at a reasonably natural pace. The result was a large amount of data (nearly 350,000 sentences) but with an unknown number of errors, and more importantly, at unknown times in the audio recordings. In order to perform acoustic-phonetic measurement and experiments with this otherwise well-documented data, we undertook the effort to identify and correct the errors.

3. Goals of New Annotation

A transcript that is faithful to the audio is required for quality assurance in many automated analyses of the speech signal. One of the unique aspects of Mixer-6 is that each speaker read the same sentences, allowing for direct comparisons across speakers matched at the sentence, word, or even phonetic level. However, any reading error on the part of the speaker results in an unknown deviance from the written transcript (i.e., the prompts). To preserve the integrity of the transcript, we audited the corpus using a combination of automated and manual methods to retain only the sentences for which the speaker read the sentence as prompted. In particular, automatic speech recognition techniques were used to identify potential reading errors, with those sentences audited through human listening.

Each read sentence could then be sorted into one of five categories of prompt-to-utterance alignment. The first category contains sentences that are very likely correct based on "perfect" ASR recognition. The second and third categories contain sentences that contained an upper-bounded range of error(s) according to the ASR. These were audited by a skilled listener, corrected if possible, and classified as either a faithful reading of the sentence or an incorrect reading of the sentence. The fourth category contains sentences with more ASR-detected errors than our threshold permitted; these were left untranscribed. The fifth category contains sentences opportunistically recovered from category four during listening. Each of these categories will be explained in more detail following a brief description of the speech recognition process.

4. Auditing and Annotation: Methods

The time-aligned transcription of the speech signal was generated in two phases of ASR. In the first phase, the audio of the Mixer-6 corpus was decoded using an HTK system trained on the Wall Street Journal Corpus (Paul and Baker, 1992) to obtain a transcript of the speech. A reference transcript, or list of prompts, was compared to the derived ASR transcript using the sclite tool from the NIST SCTK scoring package (NIST, 2009). Any sentence with less than 100% confidence or accuracy was audited through human listening. 845 sessions were audited using this process. Approximately 700 of the audited sessions were then used to train a new acoustic model using the Kaldi toolkit (Povey et al., 2011). The final system was a single-pass time delay neural network (TDNN) system, developed from a development set of 34 sessions (~6 hours of audio) and a training set of 683 sessions (~120 hours of speech).

The decoded transcript was again scored with sclite and aligned with the reference transcript, resulting in a confidence score for each word, and a count of errors for each given sentence (insertions, deletions, and substitutions). To conserve auditing and annotation resources, a threshold was set before human listening: if a sentence was under five words, no errors were permitted; for sentences over five words, a maximum of one error was permitted. Another ASR recognition pass was then completed on only the unfiltered sentences. After this process, any sentence still containing ASR-predicted errors or with less than 100% on the sclite confidence measure was audited through human listening.

Sentences with perfect accuracy and 100% confidence were accepted without additional auditing. While false acceptances on the part of the ASR system remain in the transcript, we have some reason to believe that the evaluation system was conservative in sentences determined correct, as the number of sentences in category II (ASR-reject, human-accept) was quite high relative to the number of sentences in category III (ASR-reject, human-reject; see section 5 for details).

During listening, the auditor annotated the signal when necessary with a set of defined remarks indicating how the recording was unusual, or any manual adjustment to the ASR output. These are: reading errors ('e'), initial pronoun reduction ('r'), non-speech vocalization ('l'), environment noise ('n'), static or channel noise ('s'), manually adjusted time alignments ('m'), or manual insertion of a sentence that had been excluded by ASR ('i'). Annotations were not mutually exclusive, so many sentences have more than one. In cases with no recognizable error, no annotation was made. These annotations will be in the publicly released corpus for reference and further analysis.

Reading errors included repetition, deletion, insertion, or modification of word fragments, words or phrases. In many cases, the speaker may have produced an error early in the utterance, but corrected him-/herself, repeating the sentence again. Whenever the sentence was re-read as a whole, the start and end times of the sentence were simply adjusted and the sentence was deemed correct. All other reading errors were marked as containing an error ('e') and excluded from the final transcript. Sentence-initial pronoun reductions were not considered errors as they were considered natural readings of the sentence. These were typically cases where initial 'it's' reduced to [ts]. These were generally retained in the final corpus, but annotated with a remark for reduction ('r').

Various types of noise also resulted in a sub-par ASR evaluation. These were non-speech vocalizations such as laughs, coughs, sniffs, etc.; environment noise such as sirens or a microphone bump; and finally, static or channel noise which included electrical hums or spikes. Although originally specified as separate categories, environment noise and channel noise were often collapsed into one category and typically annotated with ‘n’. Similar to the reading error corrections, if the time alignments could be adjusted to exclude the affected signal, this was performed. Otherwise, each case was documented, and it was left to the auditor's judgment as to whether the sentence should be excluded. This was evaluated on whether the speech signal conveyed a faithful and recognizable reading of the prompt.

The final type of annotation indicated any manual corrections to the ASR transcriptions, such as adjustment of the time alignment to exclude unwanted noise or speech, or include necessary parts of the speech signal ('m'), as well as any manual addition of a sentence to the transcript ('i'). Annotations were not mutually exclusive, so many sentences are labeled with more than one note. If the sentence was deemed correct without any adjustment by the human listener, no annotation was assigned.

Many sentences were excluded from the transcription due to our filtering criteria on the ASR output. The ASR recognition “failed” on a number of utterances, perhaps due to a particular speaker's voice or asynchrony of transcript and audio from some combination of speaker and ASR error. Some of these false rejections were recovered during listening by the auditor noticing and transcribing the sentence manually. However, separating all false from correct rejections would yield diminishing returns. Future revisions of this corpus may recover some of this data.

5. Annotation: Results

In total, 1412 sessions were audited, resulting in clean sessions for 549 unique speakers. Of these, 389 speakers completed three sessions, 85 completed two sessions, and 75 completed one session.

As shown in Table 1, a total of roughly 325,500 sentences cleared our auditing process. Up to 338 sentences appeared in each session transcript, with a median of 231 sentences. About 315,000 sentences achieved 100% accuracy in the ASR evaluation, and around 9500 sentences were recovered after listening. An additional 1000 sentences were manually inserted during listening.

Type	Description	Count
I	ASR-accept	315,000
II	ASR-reject, human-accept	9500
III	ASR-reject, human-reject	3500
IV	ASR-reject, unrecovered	~6000
V	ASR-reject, manually inserted	1000
Total Recovered Sentences		325,500
Estimated Total Sentences		~335,000

Table 1. Total number of accepted and rejected sentences according to audit type. Shaded cells reflect sentences excluded from the final corpus.

Altogether, around 13,000 sentences were passed on from the ASR system for human auditing. Approximately 75% of these 'suspicious' sentences were admitted to the corpus after human listening (9500 sentences mentioned above). Of these sentences, 4000 sentences were recovered after adjusting the time-alignment; 1000 sentences contained initial pronoun reductions but were nonetheless admitted; 900 sentences contained static or noise and 100 contained non-speech vocalizations; the total number of annotations is greater than the total number of audited sentences as a good portion of the sentences were annotated with more than one remark. Over 60% of the admitted sentences passed human judgment without comment (6000 sentences), suggesting that the evaluation score was relatively conservative in accepting sentences.

For the remaining 25% of audited sentences, human auditors were in agreement with the ASR evaluation and rejected the sentence (3500 sentences). Roughly 85% of the removed sentences were due to reading errors with the remaining 15% removed for intrusive noise or non-speech vocalizations.

We estimated from LDC records that approximately 335,000 sentences were read in the original corpus. Around 6000 sentences were filtered from the ASR recognition using our error criteria. After auditing, we can now estimate the sentence error rate at around 3% for the original corpus. However, in addition to knowing the error rate, we have also identified the locations of each of these errors such that they can be avoided if necessary.

The new release of the corpus will include the annotations of the ASR-generated output in addition to the final cleaned session transcripts.

Status	Comment	Count
Accepted	Adjusted time-alignment	4000
Accepted	Initial pronoun reduction	1000
Accepted	Noise	900
Accepted	Non-speech vocalizations	100
Accepted	No comment	6000
Rejected	Reading error	3000
Rejected	Other	500

Table 2. Number of sentences by comment type and acceptance status.

6. Discussion

6.1 Corpus Potential and Application

Mixer-6 provides a large corpus of transcribed American English speech for acoustic-phonetic analysis, with substantial data per participant. This is ideal for studying phonetic variability not only in a population, but also within and across individuals. Phonetic analyses depend on reliable segmentation of the speech signal, and in many cases, large data sets for sound results and real world applications. The audited transcripts particularly increase the value of Mixer-6 for phonetic study and forensic applications.

One example from our laboratory is the study of speaker variability and systematicity in the realization of speech sounds. Using a subset of the Mixer-6 read speech, Chodroff et al. (2015) examined word-initial stop consonant voice onset time (VOT) from approximately 130 speakers. The cleaned transcripts were used to produce an automatic forced alignment of the speech signal and identify the location of all relevant stop consonants. The final analysis comprised just over 68,000 stop consonants with an average of 531 stop consonants per person. Statistical analysis showed substantial talker variability but also structured variability in speaker-mean VOTs, with strong correlations across stop consonant categories (see Chodroff et al., 2015 for detail). The results and their interpretation were particularly facilitated by the quality and quantity of data not only within the entire speech corpus, but also per speaker.

In a different direction, we have also had success using Mixer-6 read speech data to evaluate text-dependent speaker identification based on pronunciation, using ASR-style recognition technology. The concept is the same as that, for example, in Andrews et al. (2001), where machine transcription with sub-phonemic (i.e., phone-like) tokens was shown to automatically capture speaker-specific information that is orthogonal to the usual speaker recognition features. A sub-phoneme-level tokenization can be extracted from speech decoding in a Kaldi ASR system. The sub-phonemic units can be used to identify different pronunciations of the same word, and thereby differentiate speakers. The validity of the transcript is fundamental: the analysis relies on the assumption that the spoken text is held constant, so any mistakes would result in different tokenizations with a potential for obscured or misleading speaker effects.

A comparable quantity of speech data is difficult to achieve in laboratory settings; furthermore, laboratory speech, while beneficial for many scientific purposes, is nonetheless farther removed from more naturally occurring speech. Relatedly, other speech corpora may provide comparable numbers of speakers (e.g. TIMIT; Garofolo et al., 1993), but fewer data points per speaker; transcribed spontaneous speech (e.g., Buckeye Corpus; Pitt et al., 2005), but fewer speakers; or, transcribed spontaneous speech, but again fewer number of data points per speaker (e.g., transcribed portion of the Switchboard corpus; Godfrey et al., 1992; Greenberg et al., 1996). The Wall Street Journal Corpus may fulfill similar criteria with a large quantity of transcribed data from many speakers (Paul and Baker, 1992), but as the same set of sentences was always used, the lexical and prosodic factors of the Mixer-6 corpus are relatively matched across talkers. This quality is especially beneficial for many questions of phonetic, forensic, and engineering research, a few of which are discussed above.

6.2 Methodology Potential and Application

The methodology employed can potentially be applied to auditing other speech corpora. Given that the task at hand was read speech, many researchers may make the simplifying assumption that the text was read correctly in all cases. As mentioned before, however, a greater degree of validity is required for many speech analyses. Human auditors may comb through the many hours of data in order to verify the read speech, but this task would be quite unwieldy. In contrast, a pure automatic approach that aligns the transcript and excludes imperfect alignments may be unnecessarily conservative. The automatic approach may retain sufficient data for many research questions, but the combination of automatic and manual methods allows for better data retention as well as an analysis of the reading and recording errors.

In the case of Mixer-6, the sentence prompts served as a strict transcript from which we could detect deviations in the speech output after aligning the transcript to the audio. This alignment and detection process easily extends to other large corpora of read speech. It could also extend to cases in which the speech may be highly constrained by a script, but with some deviations expected, such as in broadcast news. In high quality speech, such as broadcast news, even an automatic decoding of the speech greatly facilitates human transcription (e.g., Bazillon et al., 2008). The additional component of automatic error detection, or evaluation of the alignment goodness (via sclite), identifies the location of deviant segments. The human auditor can directly target these hypothesized errors for manual transcription, minimizing time spent listening to correctly aligned audio, and in the process, minimize manual effort. Overall, this methodology contributes to the further advances made in human-machine collaboration for improved speech transcription (e.g., Roy and Roy, 2009), particularly with respect to auditing the corpus transcription for use in speech research.

6.3 Future directions

Many future projects can be carried out in relation to the Mixer-6 corpus and also the methodology developed for auditing transcription. Within the read speech of Mixer-6, we would like to recover other correctly read sentences that were filtered from the output. For reasons likely due to poor ASR decoding and/or asynchrony, the derived confidence of the alignment was particularly low on many sentences: a few of these sentences were identified during human auditing and manually inserted, but many others are still untranscribed in the final output. The current configuration filters sentences based on the aggregate alignment of all words in a sentence. Relaxing the hard threshold for removing sentences would result in a greater number of sentences to audit, but also better data retention. These goals may be facilitated by outsourcing human auditing to Mechanical Turk (e.g., Marge et al., 2010). Additionally, the methodology of alignment and evaluation for error isolation can be applied to other read or highly scripted speech corpora to ensure greater validity of the transcript.

7. Conclusion

The present paper describes the auditing process used to improve the validity of the Mixer-6 read speech transcripts. Automatic speech recognition combined with human listening enabled us to audit an estimated 335,000 recorded sentences, and verify with reasonable confidence appropriate readings for 325,500 sentences. The improved validity of the transcripts enhances the inherent value of Mixer-6 for use in acoustic phonetic and SR technology research.

8. Acknowledgments

We would like to thank Alessandra Golden for substantial contribution to the auditing process, with additional thanks to Elsheba Abraham, Chloe Haviland, Spandana Mandaloju, and Benjamin Wang. We also thank Wade Shen, Sharon Tam, and Colin Wilson for contributions to data processing. Finally, we acknowledge the USSS-DHS Forensic Services Division and JHU Science of Learning Institute for supporting this project.

9. References

Andrews, W., Kohler, M., Campbell, J., Godfrey, J. 2001. Phonetic, Idiolectal, and Acoustic Speaker Recognition. In *Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop*: Chania, Crete, Greece.

Bazillon, T., Estève, Y., Luzzati, D. 2008. Manual vs assisted transcription of prepared and spontaneous speech. In *Proceedings of the 6th International Conference on Language Resources and Evaluation:* Marrakesh, Morocco.

Brandschain, L., Graff, D., Walker, K. 2013. Mixer 6 Speech LDC2013S03. Hard Drive. Philadelphia: Linguistic Data Consortium.

Chodroff, E., Godfrey, J., Khudanpur, S., Wilson, C. 2015. Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. In *Proceedings of the 18th International Congress on Phonetic Sciences*: Glasgow, UK.

Cieri, C., Campbell, J. P., Nakasone, H., Miller, D., Walker, K. 2004. The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*: Lisbon, Portugal.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, *93*.

Godfrey, J.J., Holliman, E.C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517–520.

Greenberg, S., Hollenback, J., Ellis, D. 1996. The Switchboard transcription project. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proceedings of the 4th International Conference on Spoken Language*, Philadelphia, PA, USA, pp. S24–S27.

Marge, M., Banerjee, A., Rudnicky, A. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5270-5273.

NIST. 2009. Speech recognition scoring toolkit (SCTK). Version 2.4.0. http://www.nist.gov/speech/tools.

Paul, D. B., Baker, J. M. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing:* Banff, Alberta, Canada, pp. 899-902.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., Raymond, W. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89-95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop*.

Roy, B. and Roy, D. (2009). Fast Transcription of Unstructured Audio Recordings. In *Proceedings of the 10th INTERSPEECH Conference*: Brighton, UK.