Interspeech 2018
2-6 September 2018, Hyderabad

LOCUST - Longitudinal Corpus and Toolset for Speaker Verification

Evgeny Dmitriev[1], *Yulia Kim*[2], *Anastasia Matveeva*[2], *Claude Montacié*[1], *Yannick Boulard*[4], *Yadviga Sinyavskaya*[3], *Yulia Zhukova*[2], *Adam Zarazinski*[5], *Egor Akhanov*[5], *Ilya Viksnin*[2], *Andrei Shlykov*[2], *Maria Usova*[2]

[1]Sorbonne University, France
[2]ITMO University, Russia
[3]Higher School of Economics, Russia
[4]SynAck Conseil, France
[5]Inca Digital Securities, USA

Abstract

In this paper, we set forth a new longitudinal corpus and a toolset in an effort to address the influence of voice-aging on speaker verification.

We have examined previous longitudinal research of age-related voice changes as well as its applicability to real world use cases. Our findings reveal that scientists have treated age-related voice changes as a hindrance instead of leveraging it to the advantage of the identity validator. Additionally, we found a significant dearth of publicly available corpora related to both the time span of and the number of participants in audio recordings. We also identified a significant bias toward the development of speaker recognition technologies applicable to government surveillance systems compared to speaker verification systems used in civilian IT security systems.

To solve the aforementioned issues, we built an open project with the largest publicly available longitudinal speaker database, which includes 229 speakers with an average talking time exceeding 15 hours spanning across an average of 21 years per speaker. We assembled, cleaned, and normalized audio recordings and developed software tools for speech features extractions, all of which we are releasing to the public domain.

Index Terms: longitudinal corpus, speaker verification.

1. Introduction

Speaker verification systems have recently grown in popularity as one of the authentication factors that supplements other authentication methods in call centers. Financial institutions have been particularly in need of robust multi-factor authentication methods as they move their interactions with clients online and close local branches [1]. Rapidly improving speaker verification systems are trying to fill this growing niche, and commercial offers are growing in number. Under a closer look, however, it is apparent that the voice biometrics industry simply repurposes government-funded speaker recognition "black box" technologies for civilian use. Although researchers often use speaker recognition and verification as interchangeable terms [2], there is a clear distinction between the two when considering task complexity and expected error rate. Speaker verification validates an identity claim by comparing two voice models. Speaker recognition, however, involves recognizing the voice of a speaker among all speakers in a database, which requires more computing power and produces a higher error rate. It is therefore a more difficult task. While it is hard to imagine the need for speaker recognition capabilities in civilian use cases, there is a clear bias toward speaker recognition technologies compared to speaker verification technologies.

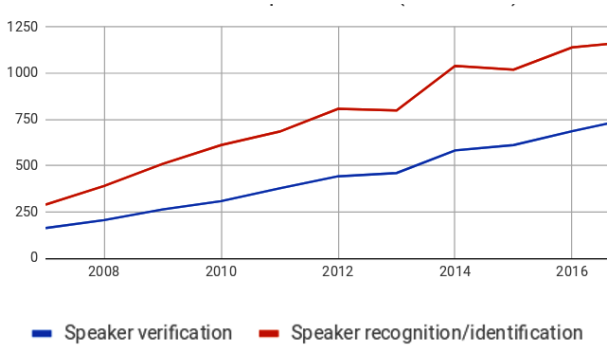

Figure 1: *Speaker verification vs speaker recognition/identification terms in INTERSPEECH publications from 2007 to 2017. Recovered using Google scholar search engine.*

By focusing on speaker recognition challenges specific to surveillance technologies, researchers often overlook certain aspects of voice biometrics that are a major obstacle for civilian authentication systems. Voice aging is one of them.

While studying voice aging in speaker verification systems, we often came across research that use paywall datasets, do not share analytical tools, or impose restrictive licensing. This paper and its associated resources are an attempt to lay the groundwork for advances in longitudinal speaker verification systems as well as to build an open community around such systems. We make all our resources available under a CC-BY-NC license to encourage other researchers and companies to contribute to the project.

2. State of Art

Voice biometrics research is one the many areas where various governments, particularly the United States, have been dictating the direction and the extent to which technology can be applied to civilian use cases. As a result, many aspects of voice biometrics research have been overlooked or misunderstood. Longitudinal speaker verification is one of these aspects. When approaching this problem, most researchers simply point out that beyond 3-5 years, the effects of aging on voice significantly increase Equal Error Rate (EER) and suggest ways to compensate for it [3]. Those

10.21437/Interspeech.2018-2412

specializing in longitudinal speaker verification tend to use calibration and compensation terms, giving age-related voice changes an unnecessarily negative connotation [4].

Considering age-related voice variations as a problem in speaker verification is a questionable security assumption. These variations are a natural process observed in all humans and follows a set of similar patterns. Researchers often overlook the fact that finding predictable voice change patterns can not only reduce EER in real world speaker verification systems, but also improve their robustness against synthesis and repeat attacks [5].

Researchers also assume that the attacker not only has similar-or-above analytical tools and computing power, but also the same amount of data to work with. In the real world, an attacker, even backed by state actors with massive data collection programs, does not have as much clean and well structured longitudinal data as narrow-scope voice authentication systems. Testimonials from NSA whistleblowers William Binney and Thomas Drake also indicate that increasing the number of individuals monitored has decreased the quality of source material and resulted in lowered capabilities to identify specific individuals [6]. It is clear that in the majority of civil use cases, the validator has a significant advantage in terms of collected audio samples, and therefore, is able to build age-aware speaker verification models that are robust to various attack vectors.

It should be noted that speech parameters used for verification can be affected by a multitude of factors that may include, but are not limited to: a speaker's sex, physiological and psychological conditions, hearing loss, disease, medications, smoking, and more [7, 8]. Although researchers point out the difficulty in separating the effects of aging from other influencing factors, they did establish a correlation between age and speech rate, sound pressure level, and fundamental and formant frequencies [8, 9, 10]. For example, a young child’s voice frequency is typically between 500Hz and 350Hz, while an average adult male voice frequency is less than 130Hz after the age of 40 and around 150Hz after the age of 80. Verification problems result from this change. A longitudinal voice change study found that voice pitch changes during 3-4 year time intervals deteriorates the performance of speaker identification by 40% and that the performance of a speaker identification system degrades by approximately 20% every 1-2 years [30].

Researchers observed decreased speaking fundamental frequency and strongly increased voice onset time with increased age [11]. Although many characteristics of voice aging are shared across individuals, there is also evidence that some age-related variations might not follow patterns in different people, but are caused by individual adjustments to physiological changes in speaker's vocal tract [10, 12].

3. Data collection

3.1. Existing longitudinal corpora

Researchers often cite corpus limitations as major obstacles in speaker recognition and verification research [3, 13]. Existing speaker verification longitudinal corpora is typically taken from one of three sources: NIST Speaker Recognition Evaluation corpora, broadcast media content, or self-built corpuses from voluntary audio recordings.

NIST Speaker Recognition Evaluation is a U.S. government funded challenge of text independent speaker recognition systems. Every two years NIST releases a corpus consisting of a large number of individuals speaking over multiple years. Although the NIST releases are some of the most sizable datasets both in terms of number of speakers and time period covered, they suffer from poor data annotation, inconsistency of recording conditions, and speaker involvement over time.

Researchers who choose radio, TV, or other broadcast media can obtain a well structured corpus of audio segments, but they often suffer from a limited number of media figures from which to source audio. Those media figures, further, are often recorded in a wide range of conditions and formats, often with other speakers talking over each other.

The most extensive studies of longitudinal speaker verification involved creating a separate corpus that includes well annotated audio recordings of individuals that span over decades. The main disadvantage of this approach is that it requires researchers to retroactively collect high-quality audio content, which often results in obtaining only a limited number of speakers or reduces the study time span. Even if this approach happens to be successful, researchers who choose to build their own corpora face scalability limitations. Authors of RedDots and TCDSA acknowledge these limitations [3, 14]. To illustrate our point, we created a comparative table of the most popular corpora used in speaker verification and voice aging studies.

Table 1: *Comparison of longitudinal audio corpora*

Corpus	Number of participants	Average time span (days)
LOCUST top speakers	229	6862
LOCUST common part	8874	731
Speaker Ageing Database [29]	18	16425
TCDSA [21]	26	15148
MARP [28]	60	1095
RedDots[22]	62	1095
NIST SRE 2010 (female part) [23]	1365	730
NIST SRE 2014 [23]	6087	730
NIST SRE 2006 [23]	504	730
TIMIT[24]	43	14
RUSBASE[25]	80	1
LLHDB[26]	40	1
CHiME[27]	34	1

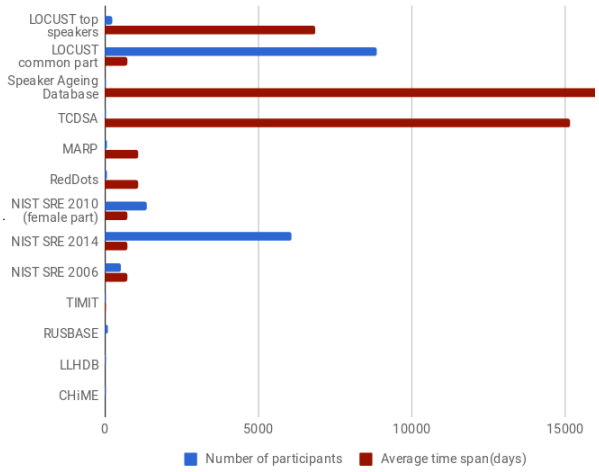

Figure 2: *Comparison of longitudinal audio corpora.*

3.2. LOCUST

3.2.1. FOIA Requests

These limitations were the impetus for us to search for a new original corpus. Our first thought was to go to the most abundant source of voice data: U.S. surveillance systems. Our idea was to use Freedom of Information Act (FOIA) requests to acquire the necessary data, which would automatically become part of the public domain if released. FOIA requires any federal agency in the United States to transmit stored information to whoever requests it. This legislation equally applies to agencies with investigative powers, such as the FBI and DEA, with some exceptions. Given the power of surveillance programs in the U.S., we assumed that their databases contain all the information necessary to build an exhaustive corpus.

A total of nine divisions within departments of Justice, Treasury, and Homeland Security were contacted to establish working relationships and assist in submitting an official FOIA request. After the first data started to flow in, however, it became apparent that the varying quality and poor annotation would have been a major obstacle in conducting any meaningful scientific research. To solve the aforementioned quality issues, we decided to focus on better structured data that is already in public domain in the form of evidence presented in federal courts. That approach also turned out be futile due to the narrow coverage and technological limitations of PACER, the central system that collects and stores court case materials.

3.2.2. US Supreme Court data

During our search for court case data, we came across a well structured dataset called Oyez. It is a collection of audio recordings from the U.S. Supreme Court, accompanied by well structured transcripts and detailed speaker information.

Table 2: *Statistics on original Oyez dataset*

Cases with transcripts and with audio	7755
Audio files	7951
Speakers	8973
Total duration	7966 hours
Total time span	62 years

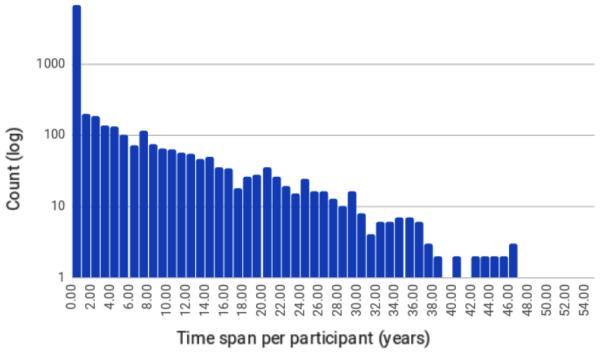

Figure 3: *Time span distribution across all speakers (log scale).*

As illustrated above, this resource, which is mostly used by legal scholars, contains a vast amount of audio data. More importantly, due to the nature of U.S. legal systems, many speakers appear in transcripts over multiple decades. This is especially true for Supreme Court judges who get appointed for life, tend not to retire, and participate in court proceedings until their death. They are also recorded in similar conditions, using similar - if not the same - equipment over time, minimizing channel effects. The nature of court proceedings also contributed to the quality of the dataset, minimizing the cases where multiple speakers talk at the same time or change their position in relation to the microphone. A homogeneous recording environment and minimal speaker overlap makes this corpus extremely relevant for studying age-related voice changes.

To build the LOCUST corpus, we created software tools that allowed us to automatically download all audio files, split them into segments based on speakers found in corresponding transcripts, and conduct various corpus cleaning procedures. During the cleaning stage, we came across and fixed problems such as low quality audio, timestamp inconsistency, missing data, and appearance of speakers who, according to their Wikipedia pages, were deceased. All of the collected data are valuable for further research, but in order to conduct voice aging related experiments, we focused on cleaning and normalizing data from speakers who appear in at least ten different sessions.

Table 3: *Statistics on top 229 speakers of LOCUST corpus (10 or more recording sessions)*

Total number of speakers	231
Males	206
Females	23
Average total speaking time	14.82 hours
Average number of appearances	231
Average time span (longitude)	18.9 years

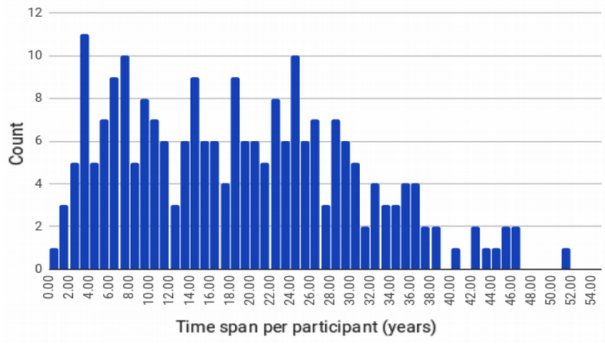

Figure 4: *Time span distribution across top 229 speakers (10 or more recording sessions).*

3.2.3. Limitations

It is also worth noting several limitations we came across while putting together this dataset. Some of the early audio recordings were made in the 1950s and the 1960s, and are of low quality and suffer from low volume and static noise. Both were relatively easy to fix during the audio normalization stage, which included adjusting average amplitude to -20 dbFS, but still suggest possible scarcity of useful audio features that can be extracted from these recordings. Also, all audio files were standardized by setting the sample rate to 16 kHz. In addition to low quality audio, statements with significant speaker overlap were removed. We are continuing our work cleaning out audio files and building tools to automate this process.

For short speech statements, we found sub-second timestamp inconsistencies between the transcript and the audio. This was mitigated by appending an additional second of audio to every speech statement we extracted. Additionally, we noticed a few cases when the speaker is clearly not the one identified in the transcript. For the moment, we removed these statements manually, but we are planning to do further corpus cleaning by using a speech verification system and double-checking any statements beyond a certain probability threshold.

Finally, Supreme Court proceeding participants are mostly well established judges and lawyers and are mostly older male U.S.-native speakers. In our selection, only 23 are female, amounting to just over 10% of the total number of speakers.

4. Acoustic feature extraction

Fundamental scientific work in speaker recognition can be easily applied to text-independent longitudinal speaker verification by selecting acoustic parameters that are robust to inter-session variability. One of the more popular methods of robust feature extraction of speech data are Mel-Frequency Cepstral Coefficients (MFCC) and Relative Spectral Transform - Perceptual Linear Prediction [15 -18]. We have built python scripts that allowed us to extract speech features from each audio segment and create corresponding MFCC and PLP-RASTA files. While performing feature extraction, the audio files were split into frames of 25 ms with an overlapping offset of 15 ms.

Denoising and normalization are also considered important parts of a speaker verification system. We have already performed average amplitude normalization while cleaning the corpus, but expect denoising to be done during the analysis stage. Thanks to a vast number of audio segments, we aim to build robust noise models based not only on the statements of the speakers we are studying, but also on others who participated in the same court session, but were not included in the final selection of 229 speakers due to their rare appearances.

5. Further research

Speaker recognition is a sensitive subject across intelligence agencies. It is our strong belief that the way forward for closely related speaker verification is an open research process from start to finish. Researchers seem to be in an arms race of speaker recognition technologies by creating competing systems to score better in challenges that lack transparency and have questionable applicability to civilian systems. Our approach will instead be based on an open collaboration with the entire industry. We are planning to expand the dataset using C-SPAN videos, perform deep analysis based on collected data, and create open source prototypes of a speaker verification system.

We also hope that the data and accompanying tools might be useful to other branches of speech studies. Although we have thus far focused on acoustic speech features, we kept all of the original transcripts, which would enable scientists to perform text-dependent analysis of our corpus. The court proceedings present a unique opportunity to study other subjects as well, such as emotion recognition [19] or lie detection. Both are relatively easy to annotate based on the transcripts and associated metadata on court case outcomes.

We would like to extend the invitation to other researchers and companies to join our GitLab project [20]. With voice, we have been given a powerful forward-secrecy signing mechanism with private keys that are hard to steal. We must not ignore it.

6. Conclusions

We found that existing datasets, assembled to advance speaker recognition technology, which is most often used by intelligence agencies, often have limited utility for developing civilian speaker verification systems. To solve this, we built an open project with the largest publicly available longitudinal speaker database, which includes 229 speakers with an average talking time exceeding 15 hours spanning across an average of 21 years per speaker. In addition to this, we have collected additional audio recordings of over 8,000 people, recorded at the same time, in the same conditions, and using the same equipment. We believe this will significantly simplify the task of building audio channel models and denoising the targeted speaker audio. Age-related voice changes that can easily be observed in our corpus will enable the creation of more robust voice verification methods. This can help the early adopters of voice authentication systems, such as call centers and penitentiary systems, that are beginning to face voice aging.

We propose an open project framework for collection, processing, analyzing, and comparing datasets. At the moment our consortium consists of scientists from three universities and two IT security companies working together under a common open GitLab project and sharing all internally developed tools under a permissionless license.

7. References

[1] S. Gold, “Financial services sector puts voice biometrics at heart of fraud battle,” Biometric Technology Today, pp. 5-9, Elsevier, 2014.

[2] M. Todisco, H. Delgado, and N. W. Evans: “Articulation Rate Filtering of CQCC Features for Automatic Speaker Verification,” INTERSPEECH, 2016.

[3] F. Kelly and J. H. Hansen, “Score-aging calibration for speaker verification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(12), pp. 2414-2424 , 2016.

[4] F. Kelly, N. Brümmer, and N. Harte. Eigenageing compensation for speaker verification. In INTERSPEECH, 2013.

[5] T. Kinnunen, Z. Wu, K. Lee, and F. Sedlak, “Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech,” Acoustics, Speech and Signal Processing (ICASSP), pp 4401 - 4404, 2012.

[6] M. V. Hayden, “Playing to the Edge: American Intelligence in the Age of Terror,” Penguin Press, 2016.

[7] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in Acoustic Characteristics of the Voice Across the Life Span: Measures From Individuals 4–93 Years of Age," Journal of Speech, Language, and Hearing Research, August 2011, Vol. 54, 1011-1021, 2010.

[8] S. Schötz, “Acoustic Analysis of Adult Speaker Age” In: Müller C. (eds) Speaker Classification I. Lecture Notes in Computer Science, vol 4343, 2007.

[9] J. Harrington, S. Palethorpe, C. I. Watson, : "Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers," INTERSPEECH, 2007.

[10] P. Torre and J. A. Barlow, “Age-related changes in acoustic characteristics of adult speech,” Journal of Communication Disorders 42(5):324-33, 2009.

[11] W. Decoster and F. Debruyne, "Longitudinal voice changes: facts and interpretation," J Voice. 2000 Jun;14(2):184-93, 2000.

[12] S. E. Linville and J. Rens, "Vocal tract resonance analysis of aging voice using long-term average spectra," J Voice. 2001 Sep;15(3):323-30, 2001.

[13] D. E. Sturim, P. A. Torres-Carrasquillo, J. P. Campbell, “Corpora for the Evaluation of Robust Speaker Recognition Systems,” INTERSPEECH, 2016.

[14] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, “ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” INTERSPEECH, 2015.

[15] S. J. Chaudhari, R. M. Kagalkar, “Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition,” International Journal of Computer Applications, V. 117 - No. 17, 2015.

[16] J. V. Psutka and L. Müller, “Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task,” INTERSPEECH, 2001.

[17] A. K. Lee, A. Larcher, C. H. You, B. Ma, and H. Li, “Multi-session PLDA Scoring of I-vector for Partially Open-Set Speaker Detection,” INTERSPEECH, 2013.

[18] H. Erokyar, “Age and Gender Recognition for Speech Applications based on Support Vector Machines,” 2015.

[19] C. Montacié, M-J. Caraty, “High-level speech event analysis for cognitive load classification,” INTERSPEECH, 2014.

[20] LOCUST Project https://gitlab.com/evgenyd/LOCUST/

[21] TCDSA database kellyfp@tcd.ie

[22] RedDots database http://goo.gl/forms/Dpk3OiJkWV/

[23] NIST SRE database https://www.nist.gov/itl/iad/mig/speaker-recognition

[24] TIMIT database https://goo.gl/l0sPwz

[25] RusBase http://kingline.speechocean.com/exchange.php?id=4166&act=view

[26] LLHDB https://ucla.box.com/s/w9rfmbohgomrpms9buwa9e40gobr4jlm

[27] CHiME database http://spandh.dcs.shef.ac.uk/chime_challenge/data.html

[28] A. D. Lawson, A. R. Stauffer, E. J. Cupples, W. P. Bray, J.J. Grieco. “The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings,” INTERSPEECH 2009.

[29] F. Kelly, A. Drygajlo, N. Harte. “Speaker Verification with Long-Term Ageing Data,” 5th IAPR International Conference on Biometrics, 2012.

[30] Y. Matveev: "The Problem of Voice Template Aging in Speaker Recognition Systems," Speech and Computer, SPECOM 2013, Lecture Notes in Computer Science, vol 8113, 2013.