The RedDots Data Collection for Speaker Recognition

Kong Aik Lee[1], *Anthony Larcher*[2], *Guangsen Wang*[1], *Patrick Kenny*[3], *Niko Brümmer*[4], *David van Leeuwen*[5], *Hagai Aronowitz*[6], *Marcel Kockmann*[7], *Carlos Vaquero*[8], *Bin Ma*[1], *Haizhou Li*[1],*Themos Stafylakis*[3], *Jahangir Alam*[3], *Albert Swart*[4], *Javier Perez*[6]

[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Université du Mans - LIUM, France
[3]Centre de Recherche Informatique de Montreal (CRIM), Quebec, Canada
[4]AGNITIO Research, Somerset West, South Africa
[5]NovoLanguage, The Netherlands
[6]IBM Research – Haifa, Israel
[7]VoiceTrust, Ontario, Canada
[8]AGNITIO S. L., Madrid, Spain

Abstract

This paper describes data collection efforts conducted as part of the RedDots project which is dedicated to the study of speaker recognition under conditions where test utterances are of *short duration* and of *variable phonetic content*. At the current stage, we focus on English speakers, both native and non-native, recruited worldwide. This is made possible through the use of a recording front-end consisting of an application running on mobile devices communicating with a centralized web server at the back-end. Speech recordings are collected by having speakers read text prompts displayed on the screen of the mobile devices. We aim to collect a large number of sessions from each speaker over a long time span, typically one session per week over a one year period. The corpus is expected to include rich inter-speaker and intra-speaker variations, both intrinsic and extrinsic (that is, due to recording channel and acoustic environment).
Index Terms: speaker recognition, crowd sourcing, corpus collection

1. Introduction

The RedDots project aims to collect speech data over mobile devices primarily for the development and evaluation of automatic speaker recognition systems. Mobile devices, in the form of smartphones and tablet computers, provide tools and connections that allow people to access and share information. They have shown great potential as dominant user access points to the Internet and cloud services, and as sensory inputs for smart cities [1]. They can be used at any location, be it urban with Internet connectivity (office, home, public transport) or rural areas where Internet connectivity may be absent. Studies show that 80% of global Internet access will take place through mobile devices by 2016 [2]. Due to their small form factor, speech input has proved to be an attractive alternative to conventional text-based input via touch screen or an on-screen keypad. For instance, Google voice search [3] has shown to be successful. Voice authentication for mobile application is another example [4].

With the advances of mobile technology and ever increasing computational power, it may be supposed the built-in microphone in mobile devices would be of high quality. This may not be entirely true. To squeeze more components onto small devices, the space available for the microphone has been greatly reduced [5]. Also, the signal conditioning and enhancement recorded speech is subject to generally differs from one manufacturer to the next. Mobile devices are often used in varying environments, ranging from private offices and quiet meeting rooms to locations with noisy surroundings such as crowded cafeterias or streets. The need to compensate for variability in recording channels and environment poses a multitude of challenges to speaker recognition technology. In addition to speaker-extrinsic variation, the RedDots collection was designed to capture speaker-intrinsic variation due to speaker physical condition (e.g., flu, sore throat). This is accomplished by having speakers participate in a large number of recording sessions spaced at regular intervals over long time span. One target scenario is to collect one session per week over a one year period for each speaker.

The collection of speech corpora is typically carried out by either requesting the speakers to be onsite [6], [7] or remotely through telephone calls [8], [9], [10], [11]. The former has the benefit of a controlled environment, where the channel and acoustic environment conditions can be kept consistent across speakers. The latter allows speakers to record speech from wherever they happen to be located, which has the benefit of a potentially wider population with greater diversity. Recently, we have seen a rising trend of collecting speech data remotely using the Internet as it becomes more widely accessible [12], [13], [14]. In [12], [13] speech data were collected via a web-based interface in the presence of Internet connectivity. A slightly different methodology was reported in [14], where a dedicated mobile application (app) was developed for data collection. Similar to [14], we use a mobile app as the recording front-end. Speakers record their voices offline and later on upload the recordings to an Apache web server when Internet connection is available. This has the benefit of enabling data to be collected in the field in the absence of a persistent Internet connection.

This paper describes the database development efforts as part of the on-going RedDots project with collaboration from multiple sites. The effort was initiated as a follow-up to a special session [15] during INTERSPEECH 2014. The major motivation is to collect a speech corpus for the development and evaluation of speaker recognition system targeted for fixed

phrase, free speech and text-prompted input modes. For the evaluation of text-dependent speaker recognizers, we need a corpus having all speakers reading a similar set of sentences [7] – one characteristic that makes the design of text-dependent corpora different from text-independent corpora. The data collection procedure and protocol were designed to meet a multitude of use cases. The dataset consists of four parts of increasing degree of lexical variability. It is also our intention to make the database useful for the study of inter and intra speaker variability modeling with short utterances.

2. The RedDots Dataset

2.1. Design specification

Contrary to previous work [7], the current dataset was designed to have a relatively small number of utterances per session – speakers participate in a larger number of sessions of shorter duration instead. The current targeted scenario is to collect 52 sessions per speaker, that is, one session every week for a year. The reason for recording more sessions is to obtain sufficient coverage of intra-speaker variability over time. This strategy also fits in better with the usage of mobile applications, where users tend to switch between applications given the multi-purpose nature of the devices. Each session is limited to two minutes, with individual utterances being of 3 seconds duration on average. Allowing for the transition time between utterances, the number of sentences is set to 24 for each session. The composition of sentences used for a recording session is shown in Table I.

As shown in Table 1, the sentence set is designed to consist of four parts of increasing lexical variability. Part I consists of ten sentences common to all speakers; Part II consists of ten sentences unique to each speaker; Part III consists of two free-choice sentences chosen by the speaker; while Part IV consists of free-text sentences that are unique across sessions. The first ten common sentences are pronounced by all speakers across all sessions. Lexical variability is thus limited to ten sentences. The ten unique sentences of Part II and the two free-choice sentences of Part III are speaker specific and they differ from one speaker to another. Thus we are sampling the lexical space at $N \times S$ points instead of N in the former case, where N and S are the number of sentences and speakers, respectively. The fully variable case happens in Part IV, where lexical content change across speakers as well as all the sessions of the same speaker. Lexical variability is folded in at each of the M sessions, so we now sample the lexical space at $N \times S \times M$ points. (We use N to denote the number of sentences in each part, though its value differs from one part to another as indicated in Table 1.)

Table 1: *Type of sentences used for each recording session.*

Sentence Type	#Sentences
Common	10
Unique	10
Free choice	2
Free text	2

2.2. Use case consideration

2.2.1. Short test utterances

The use of subspace modeling techniques [16] has greatly improved the robustness of speaker verification against variability due to channel [17], [18] and noise [19]. The main limitation of these techniques is their strong dependence on large amounts of data for background modeling. Also, it is generally conceded that text-independent speaker recognition systems perform poorly with utterances of short duration. This is primarily due to the difficulty in factoring out the influence of phonetic content.

Two approaches have shown to be promising for tackling the issue of phonetic variability in the case of short utterances. In the first approach, we consider cooperative speakers pronouncing the same sentence during enrollment and test. This is called text-dependent speaker verification [7]. Parts I, II, and III of the database were designed for this scenario. The second approach focuses on "content matching" which aims to equip text-independent speaker recognition systems with a way of matching words or phones that co-occur in enrollment and test utterances [20]. Part IV was designed for this scenario, where test utterances are of short duration and of variable phonetic content. Text-dependent and text-independent speaker recognition both stand to benefit from the progress in this area.

2.2.2. Low error rate region

In the deployment of speaker recognition system for access control, the prior probability of target trial is generally close to unity. This means the system is likely to operate at a small miss rate. For the purpose of calibration, or to measure performance, we need to have sufficient number of misses at the low miss-rate region on the detection error trade-off (DET) curve. To flesh out a bit, let consider a miss rate P_{miss} of 0.1%. In order to count at least 30 misses at the particular miss rate [21], the number of target trials N_{target} has to be larger than $30,000$ since $N_{\text{miss}} = N_{\text{target}} \times P_{\text{miss}}$. This is traditionally hard to achieve as target trials are relatively scarce. Furthermore, if calibration and evaluation are to be conducted on independent datasets, N_{target} has to be doubled. Similar arguments can be applied in the case of applications where the cost of false alarms is much greater than the cost of misses (e.g. prevention of credit card frauds). In this scenario, the area of interest is the low false alarm region.

Suppose that we have collected data from S speakers. The number of possible non-target trials, $S(S-1)/2$, grows as S^2. In other words, by making S large, the number of non-targets grows quadratically since we can match a speaker against others. This is not the case for target trials, whose number grows only linearly with S. However, if each speaker has M sessions ($M > 1$), the number of possible target trials, $M(M-1)/2$, grows quadratically with M if we consider all sessions as potential "models" (in the terminology of the NIST evaluations [22], [23]). In the current project, our approach is to collect large number of sessions per speaker over a long period of time. In addition to obtaining good coverage of intra-speaker variability over time, we will be able to manufacture a large number of target trials. Furthermore, we hope that having multiple sessions per speaker will facilitate the development of subspace methods for inter-session variability compensation.

2.2.3. Lexical coverage

The database design presented above includes ten speaker-unique sentences per speaker. These sentences facilitate investigation into scenarios where the speaker recognizer is exposed to unseen phrases at enrollment and test time. This is important because, in a practical scenario, even if all users are asked to use a common phrase, it can never be guaranteed that the true clients and impostors will actually be speaking the required text

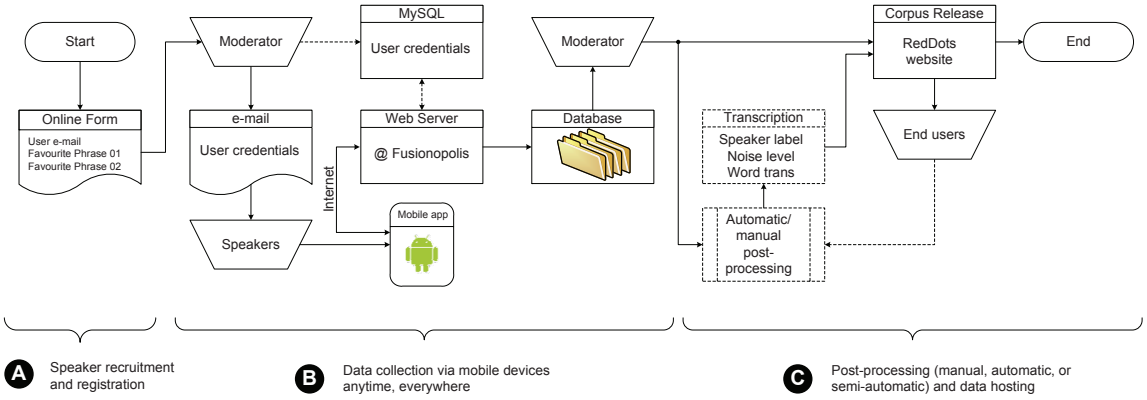

Figure 1: The acquisition protocol and process flow for data collection over mobile devices as adopted for the RedDots project

[7]. Furthermore, text-dependent speaker recognition systems generally need a passphrase rejection capability as a counter-measure to spoofing attacks and the speaker-unique sentences can be used to study this problem. The two free-choice sentence contributes as well to the same objective. In addition, we can get some insight in what people choose if they are given the choice.

3. Practical consideration

3.1. Acquisition Protocol

The flow cart in Fig. 1 illustrates the process flow from speaker recruitment and registration to voice recording and data hosting. The on-line registration form, the mobile application and the web server are the three main functional blocks of the data collection infrastructure. It is worth mentioning that the process is semi-automatic, where a moderator verifies the user inputs received via the on-line form and the voice recordings uploaded to the web server. The acquisition process flow is detailed below.

Step 1. *Speaker registration*. Prospective speakers fill in an online registration form[1]. The information collected includes e-mail address, gender, age group, country of residence, native language of the speakers, two *free-choice* sentences, and user consent to allow their voice recordings to be made available for research purpose[2].

Step 2. *User credentials and instructions*. The information collected through the registration form is verified manually and keyed-in to a database. The user credentials (a five-character user ID paired with a four-digit passcode), a user guide and the download link for the mobile app are made available to the users via e-mail. Speaker identity is tagged to the five-character user ID and e-mail address. The e-mail addresses are stored separately from other meta-data.

Step 3. *Mobile data collection anytime, everywhere*. Upon the first log-in to the mobile app, a designated list of sentences will be retrieved from the server. Once the list of sentences has been downloaded, recording can be carried out off-line. The recorded speech samples are uploaded when network connectivity is available. This feature allows users to make recordings at any indoor or outdoor locations.

[1] http://goo.gl/forms/2KmktzgVV9

[2] Users acknowledge by submitting the form

Step 4. *Data post-processing*. A first layer post-processing will be carried out to discard inappropriate recordings. It is expected that automatic or semi-automatic post-processing will be carried by research community. The transcription obtained will be made available as part of the data release.

Step 5. *Data hosting*. The data collected will be made available via an online portal. Update will be released in a regular basis.

Table 2: *Total number of sentences and their sources used for data collection*

Sentence Type	#Sentences	Source
Common	10	TIMIT
Unique	4,000	Gigawords
	3,000	News 2008
	3,000	News 2009
Free choice	2,000	User provided
Free text	104,000	Wikipedia

3.2. Corpus text material

Central to the data collection is the selection of sentences to be used as prompts. As we explained in Section 2.1, our goals with respect to lexical variability are very ambitious. Table 2 illustrates the number of distinct sentences that we need to produce in order to handle 1000 participants. The sources of the sentences that we are using are also given.

All sentences are in English. Except for the TIMIT [6] and Gigawords [24], which are standard corpora from LDC, all other sentences are sourced from the web. In particular, *News 2008* and *News 2009* are crawled from various news sites. For the free-text sentences, English Wikipedia dumps are used. [3] Before they can be used as prompts, the raw texts are segmented into sentences and normalized. Sentences are then selected so that all the phoneme realizations of each speaker can be captured weekly. More specifically, the speaker-unique and free-text sentences are selected so that all the phonemes are covered for each chunk of 10 sentences. The CMU dictionary [4] was

[3] http://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia

[4] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

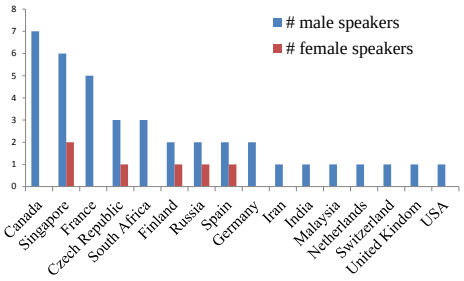

Figure 2: *Number of speakers per country as of March 16, 2015*

used to resolve words into their pronunciations. This dictionary contains more than 130K word entries covering many esoteric words that are difficult to read. We selected 20K entries with a phoneme set of 39 phonemes. The 20K word list was taken from the Wall Street Journal (WSJ) corpus. Other practical considerations when selecting the sentences are listed below:

- The total number of phonemes for each sentence is constrained to be between 15 and 25 in order to avoid sentences which are too long or too short.
- A stop list of sensitive words (e.g., religious or sexual content) is used to eliminate sentences that contain any of these words. Examples are "extremist" and "drug".

3.3. Other implementation issues

As shown in Table I, for each recording session, a speaker read a set of 24 sentences consisting of 10 common, 10 unique, 2 free-choice, and 2 free-text sentences. For each speaker, the first 22 sentences will be repeated for each session, while the 2 free-text sentences will be different from one session to the next. In our implementation, sentences in the same batch are randomly shuffled to counter the effects of learning, habituation and user fatigue that may arise as a result of repeating the same sentences in every recording session.

The RedDots dataset facilitates a longitudinal study of the impact of the aging phenomenon on text-dependent [25] and text-independent [26] [27] [28] speaker recognition with an expected time span of one year. Another benefit of the long time span is that it may be possible to capture intrinsic variation due to speakers' physical condition (e.g., flu, sore throat). To this end, a feature is made available on the mobile application which allows users to tag specific remarks regarding their recordings. On the downside, speakers tend to skip the scheduled weekly recordings. To counter this tendency, a reminder feature has been built in to the mobile application.

4. Community-driven approach

The project was rolled-out on January 29, 2015. Volunteers are recruited worldwide through personal contacts and various mailing lists. As of March 16, 2015 we have recruited 45 speakers from 16 countries. The distribution of speakers is shown in Fig. 2. It could be noted that the number of male speakers is significantly higher than female, which amount to 87%. This bias is mainly due to the fact that the current pool of volunteers are mainly those working on speech research predominantly male. In terms of age, majority of speakers belong to the $26 - 35$ age group, which amounts to 58%. These two factors pose a challenge that we have to look into as part of future work.

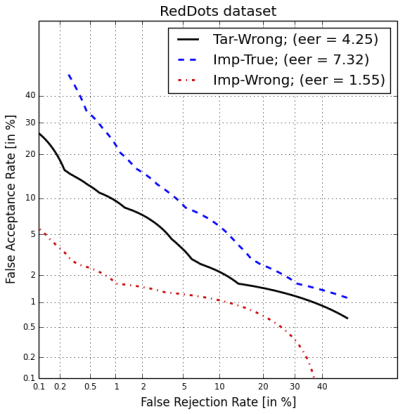

Figure 3: *DET curves on RedDot's data for 3 types of imposture*

Preliminary experiment was conducted using the male data. Only the first eight common sentences were used. This requires the speakers to complete at least two sessions (so that one session can be used for enrollment and another for testing). There are 15 male speakers with a total of 333 test utterances that fulfill this criterion. The system used for the experiment was developed based on the open source ALIZE platform [29] and the front-end processing is based on SIDEKIT[5]. An UBM with 512 mixtures was trained on data from the 157 male speakers drawn from the RSR2015 [7]. The same data was used to train a total variability matrix of rank 150 for i-vector extraction. Eventually, a PLDA model was trained with 100 eigenvoices with a full residual covariance matrix. The 3 enrolment i-vectors used for each model were averaged and scored against each test i-vector. For comparison purpose, the evaluation protocol has been kept similar to the one described in [7] for text-dependent speaker verification. For the three DET curves in Fig. 3, the case where the target speaker pronounces the correct pass-phrase is taken as the target trial. Non-target trials comprise of either target speaker pronouncing the wrong pass-phrase (*Tar-Wrong*), impostor pronouncing wrong pass-phrases (*Imp-Wrong*), or impostor pronouncing correct pass-phrases (*Imp-True*). Compared to that reported in [7], the EERs are almost doubled on the RedDots dataset. This makes the dataset challenging, and interesting.

5. Conclusions

The RedDots project is dedicated to the study of speaker recognition under conditions where test utterances are of *short duration* and varying degrees of variability in *phonetic content*. Phonetic variation is handled by having four different sentence types with increasing lexical variability. A distinguishing feature of the RedDots dataset is the high degree of inter-speaker variation which covers multiple regions worldwide. As of the time of this writing, we have recruited 45 speakers from 16 countries, with a total of 91 complete sessions. Current result shows an EER of 7.32% on text-dependent speaker verification task. The first release of the dataset is planed for the third quarter of 2015.

[5]SIDEKIT is an open-source toolkit for speaker recognition that will be released in the coming months, http://www-lium.univ-lemans.fr/sidekit/

6. References

[1] G. Cardone, L. Foschini, P. Bellavista, A. Corradi, C. Borcea, M. Talasila, and R. Curtmola, "Fostering participaction in smart cities: a geo-social crowdsensing platform," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 112–119, Jun. 2013.

[2] W. Bold and W. Davidson, "Mobile broadband: redefining internet access and empowering individuals," in *The Global Information Technology Report*, S. Dutta and B. Bilbao-Osorio, Eds. World Economic Forum, 2012.

[3] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition*, A. Neustein, Ed. Springer US, 2010, pp. 61–90.

[4] K. A. Lee, H. Li, and B. Ma, "Speaker verification makes its debut in smartphone," *SLTC Newsletter*, Feb. 2013.

[5] J. Hecht, "All smart, no phone," *IEEE Spectrum*, vol. 51, no. 10, pp. 36 – 41, Oct. 2014.

[6] V. Zue, S. Seneff, and J. R. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[7] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communucation*, vol. 60, no. 0, pp. 56 – 77, 2014.

[8] B. Wheatley and J. Picone, "Voice across America: toward robust speaker-independent speech recognition for telecommunications application," *Digital Signal Processing*, vol. 1, no. 2, pp. 45–63, 1991.

[9] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[10] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: an American English telephone speech corpus for polyphone project," in *Proc. ICASSP*, 1994, pp. 81–84.

[11] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: the Mixer 3, 4 and 5 corpora," in *Proc. INTERSPEECH*, 2007, pp. 950–953.

[12] "LibriVox - free public domain audiobooks," https://librivox.org/, accessed: 2015-02-27.

[13] "VoxForge," http://www.voxforge.org, accessed: 2015-02-27.

[14] I. Lane, A. Waibel, M. Eck, and K. Rottmann, "Tools for collecting corpora via Mechanical Turk," in *Proc. NAACL HLT*, Jun. 2010, pp. 184–187.

[15] K. A. Lee, A. Larcher, H. Aronowitz, and P. Kenny, "Is'14 special session on text-dependent speaker recognition with short utterances," Tech. Rep., Nov. 2014.

[16] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, May 2005.

[17] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[18] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[19] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition," in *Proc. IEEE ICASSP*, May 2013, pp. 6788–6791.

[20] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Proc. INTERSPEECH*, 2014, pp. 1317 – 1321.

[21] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation overview, methodology, systems, results, perspective," *Speech Comm.*, vol. 31, no. 23, pp. 225–254, Jun. 2000.

[22] National Institute of Standards and Technology, "The NIST 2008 SRE Evaluation Plan," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[23] ——, "The NIST 2010 SRE Evaluation Plan," 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/

[24] D. Graff and C. Cieri, "English gigaword ldc2003t05," Philadelphia: Linguistic Data Consortium, 2003. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2003T05

[25] W. Mistretta and K. Farrell, "Model adaptation methods for speaker verification," in *Proc. ICASSP*, 1998, pp. 113–116.

[26] Y. Lei and J. H. Hansen, "The role of age in factor analysis for speaker identificaiton," in *Proc. INTERSPEECH*, 2009, pp. 2371–2374.

[27] A. D. Lawson, A. Staufer, B. Smolenski, B. Pokines, M. Leonard, and E. Cupples, "Long term examination of intra-session and inter-session speaker variability," in *Proc. INTERSPEECH*, 2009, pp. 2899–2902.

[28] F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification," *Biometrics and ID Management (Springer)*, pp. 113–124, 2011.

[29] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. S. Mason, and J.-Y. Parfait, "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition," 2013, pp. 2768–2773.