Extended RSR2015 for text-dependent speaker verification over VHF channel

Anthony Larcher, Kong Aik Lee, Pablo L Sordo Martínez, Trung Van Nguyen, Bin Ma, Haizhou Li

Extended RSR2015 for text-dependent speaker verification over VHF channel

Anthony Larcher[1], *Kong Aik Lee* [1], *Pablo L. Sordo Martínez*[1,2], *Trung Hieu Nguyen*[1], *Bin Ma*[1], *Haizhou Li*[1]

[1] Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore,
[2] Speech and Image Research Group, Swansea University, Wales, United Kingdom

Abstract

Text-dependent speaker verification over degraded radio channel is a challenging task. To better understand the research problem, the Institute for Infocomm Research (I2R) of Singapore has collected a corpus of voice recordings transmitted over marine VHF. Built as an extension of the *RSR2015* database, the VHF-*RSR2015* consists of recordings from 300 speakers of Part I of the *RSR2015* database transmitted over VHF channel. Extending the *RSR2015* database, we would like to facilitate the study of the VHF channel effect, therefore, keeping the acoustic environment the same as that of *RSR2015*. Performance benchmark of a text-dependent speaker verification engine is given as reference on the original *RSR2015* database recorded at a sampling frequency of 16kHz, on a sub-sampled version of the same dataset at 8kHz and on the data after transmission through the VHF channel.

Index Terms: database, speaker verification, text-dependent, VHF

1. Introduction

Speaker verification is the task of confirming or declining an identity claim based on the specific characteristics of a person's voice. Speaker verification can operate in two input modes: text-independent for which the speaker is free to pronounce any speech content or text-dependent where the speaker is required to pronounce a pre-defined lexical content[1, 2, 3].

Recently, the Institute for Infocomm Research (I^2R) undertook a project with the Maritime and Port Authority of Singapore (MPA[1]) to evaluate the performance of speaker verification engines over the VHF channel. In the current scenario, ship-masters who get an accreditation from the authority are allowed to navigate within the port area without the need to engage a piloting service. The ship-master is required to contact the port controller for clearance after verifying that: 1) the caller is the ship-master with an up-to-date accreditation, 2) the caller is on board of the vessel. As an alternative to the VHF radio, cellphone is definitely better (provides a better channel) for speaker verification systems but does not assure that the ship-master is on board. Due to such operational constraints on the authentication process, text-independent speaker verification has not been considered.

The difficulty of speaker verification engines to deal with channel mismatch has been illustrated in the latest NIST-SRE evaluations [4, 5], and the RATS DARPA program [6, 7, 8] shown that speaker verification and voice activity detection (VAD) in degraded channels, such as VHF, are even more challenging [9, 10]. Additionally, both NIST-SRE and RATS tasks focus on long duration speech for both enrollment and test. It is commonly known that channel compensation does not work as well for the case of short durations.

In a survey of publicly available databases for text-dependent speaker verification, reported in [11], we showed that corpora specially designed for research in this area are very limited in terms of quantity of data and speakers variability. Moreover, specificities of the application scenarios make most of the databases application dependent such that is is extremely difficult to pool the speech resources for research or development. More troublesome, only 5 databases out of the 24 listed in [11], include data recorded with environment or channel mismatch [12, 13, 14, 15, 16]. Recently, the *RSR2015*[2] database, with 300 English speakers (153 male and 147 female) recorded on mobile devices for text-dependent speaker verification, was collected by I^2R [17, 11]. In order to specifically study the effect of marine VHF channel, the 71h of speech signal from Part I of the *RSR2015* were transmitted over the VHF radio and recorded at the far end.

In the remaining of this paper, we give a description of the Part I of the *RSR2015* database. We then describe the platform used for recording and the resulting new data. We provide an experimental benchmark on the original data recorded at a sampling rate of 16kHz, on the same data after down-sampling to 8kHz, and also after transmission over VHF. A description of the system used for this benchmark is given together with the experimental protocol. Results and analysis form the last part of this paper.

2. RSR2015 database Part I

Part I of the *RSR2015* includes 71 hours of speech data that have been collected for a scenario of text-dependent speaker verification in which a speaker has to pronounce a fix pass-phrase to be authenticated. Each of the 300 speakers was given 3 mobile devices labeled *A*, *B* and *C* that were used in sequence {*A, B, C, A, B, C, A, B, C*} to complete 9 sessions of recording. The 3 devices were chosen from 6 available options (4 smartphones and 2 tablets). During a recording session, all speakers pronounce the same set of 30 pass-phrases extracted from the TIMIT database [18], an example of which is: *"Only lawyers love millionaires."* The recording took place in a typical office environment. A push-to-talk feature was implemented into an android© application so that the recording start and stop was controlled by the user. The sentence to read was display on the screen. The participant was free to hold the device in a way (s)he feels comfortable. The audio signal was recorded through the internal microphone at a sampling frequency of 16kHz and transmitted to a server in raw PCM format (16bits per sam-

[1] http://www.mpa.gov.sg/ accessed March 28, 2014

[2] http://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications

ple with linear encoding). A SPHERE[3] header was then added to include meta-information. The *RSR2015* database include two other parts dealing with different lexical constraint, namely, short commands for control of a smart-home prototype and digit strings for a text prompted scenario [11].

3. The VHF extension

3.1. The VHF transmitting-recording platform

The Very High Frequency (VHF) radio communication complies electromagnetic waves from 30 MHz to 300 MHz [19]. The frequency range from 156 MHz to 162.025 MHz is known as *marine VHF* and is restricted to the use of maritime communications. The frequency allocated to this project is 170.375 MHz, which is slightly higher than the marine VHF.

The aim of this transmit-and-record process is to collect samples over a VHF channel in conditions similar to a real communication. The VHF being half-duplex channel technology makes the recording process complex for two reasons. First, the receiver side is never aware of when the communication start and thus, synchronization has to be addressed separately. Second, the signal on the receiver side is zero until the transmitter turns on and create a noise floor. The different signal level: *zero*, *noise floor* and *speech* may affect the VAD.

In order to address those issues regarding VHF communication particularities, we set up an automated platform to record the 80,888 transmitted speech samples of the Part I of the *RSR2015* database together with the *on* and *off* clicks. The hardware set-up consists of a transmitter at one end and a receiver at the other end. Both sides include a laptop to respectively play and record the original signal over the VHF channel, an external USB sound card (Creative Sound Blaster X-FI HD) and the VHF transceiver (Motorola GP328). The transmitting side also includes the afore mentioned PTT device, activated by a USB-controlled mechanical relay. Both transmitter and receiver are synchronized via NTP servers [4]. Figure 1

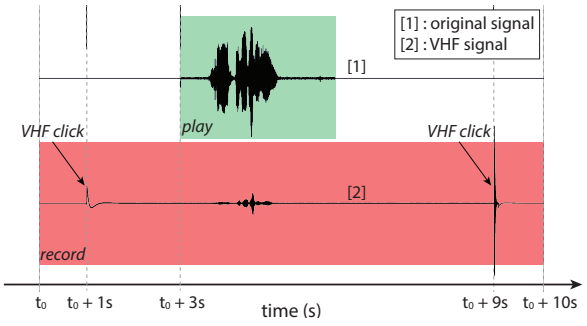

Figure 1: *Example of a waveform from the RSR2015 database before (upper panel) and after (lower panel) transmition-reception over the VHF channel. On the lower panel, clicks from the VHF tramsitter are visible before and after the speech signal.*

describes the methodology followed. At t_0 the receiver starts recording. One second later the transmitter side activates the PTT transceiver enabling VHF communication. The clean signal is played two seconds later, while the VHF channel remains opened six seconds more until the mechanical relay disables it. The recording finishes after ten seconds since it started. This methodology ensures the complete original signal is recorded in realistic conditions.

[3] http://www.itl.nist.gov/iad/mig/tools/ (accessed on March 24, 2014)

[4] http://www.pool.ntp.org/en/ (accessed on March 28, 2014)

3.2. Collected data

Transmitting a large number of speech samples over VHF in an automated manner is a challenging task that requires constant human monitoring and maintenance. Additionally, speaker verification over VHF channel is known to be challenging [8] and the performance degradation compared to the case of the original *RSR2015* is expected to be tremendous. For this reason, we decided to transmit only the Part I of the *RSR2015* database which is the less challenging of the three parts [11] but complies with the duration limitation of the operational conditions.

Transmitted data include the 80,888 speech samples from Part I of the *RSR2015* database. Due to the transmission process (Fig. 1) that requires a delay between the beginning of the recording, the activation of the VHF and the moment the speech sample is played, the recording of the original 71h of speech signal produced a total of 224.5h of audio signal.

An example of the wave form before and after transmission are shown in the upper and lower panels of Figure 1. Attenuation of the signal conjugated with the high noise floor of the VHF channel strongly affects the VAD performance. The average duration of speech per sentence on the whole Part I, detected by our VAD which includes a speech enhancement module (sec. 4.1.1), decreases from 1.78s on the original database to 0.85s after transmission over the VHF. Figure 2 shows the spectrum of the original signal down-sampled to 8kHz and the spectrum of the same sample after transmission over the VHF. The VHF channel not only limits the bandwidth of the signal but also alters the whole spectrum.

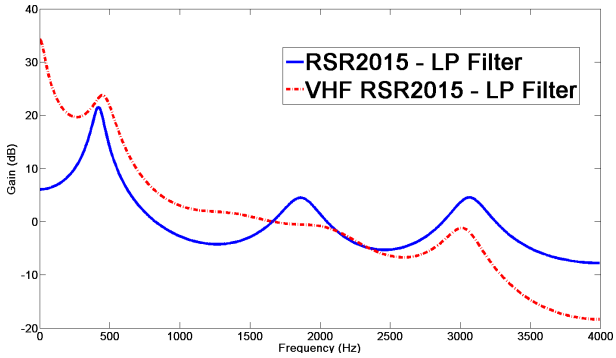

Figure 2: LPC spectra for one sentence of the *RSR2015* after down-sampling to 8kHz and transmission through VHF channel.

4. Recognition system and experimental protocol

4.1. Text-dependent speaker verification system

4.1.1. Front-end

At the front-end, the audio signal is first transformed to a sequence of frames by performing DFT on windowed-signals of 25 milliseconds in length with overlapping of 15 milliseconds.

Table 1: Performance of *HiLAM* system on the *development* and *evaluation* sets of Part I of the *RSR2015* databases and its VHF extension in terms of Equal Error Rate and minimum DCF in the sense of NIST-SRE08 (EER % / minDCF $\times 100$) for different definitions of non-target access.

	User	Target		Impostor		Male			Female		
Set	Text	correct	wrong	correct	wrong	16kHz	8kHz	VHF	16kHz	8kHz	VHF
Development	Trials	tar	non	-	-	1.00 / 0.45	0.75 / 0.27	16.10 / 8.23	0.58 / 0.25	0.44 / 0.24	12.46 / 7.24
		tar	-	non	-	1.43 / 0.63	1.71 / 0.77	10.95 / 4.95	0.97 / 0.45	1.53 / 0.76	9.01 / 4.26
		tar	-	-	non	0.20 / 0.05	0.10 / 0.03	5.79 / 2.64	0.05 / 0.02	0.02 / 0.01	4.40 / 2.06
Evaluation	Trials	tar	non	-	-	0.66 / 0.24	0.27 / 0.14	15.60 / 9.09	0.14 / 0.04	0.16 / 0.07	9.03 / 7.22
		tar	-	non	-	1.33 / 0.52	0.97 / 0.51	10.38 / 4.92	0.53 / 0.29	1.17 / 0.55	6.47 / 3.25
		tar	-	-	non	0.09 / 0.03	0.02 / 0.01	5.58 / 2.78	0.03 / 0.01	0.02 / 0.01	2.74 / 1.52

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the enrollment pass-phrase; *wrong* text means that the pass-phrase pronounced during the test is different form the enrollment one.

An energy-based speech activity detector (SAD) is then applied to select high energy frames for processing while remaining frames are discarded. Interested readers could refer to [1] for details of the speech detection algorithm. To further enhance the accuracy of frames selection, particularly in noisy environment, a standard speech enhancement module with spectral subtraction technique is employed prior to SAD. Finally, 19 MFCCs are extracted for each speech frame, with their first and second-order derivatives appended to form the feature vectors and mean variance normalization is then performed to mitigate the changes in environmental factors.

4.1.2. HiLAM engine

The HiLAM is a text-dependent verification engine based on a three-layer acoustic architecture shown on Figure 3. The speaker-dependent pass-phrase HMM is obtained through a two-step adaptation. First, a speaker-dependent GMM is adapted from a gender-dependent UBM, as follows:

$$p(o|\lambda_{gmm}) = \sum_{c=1}^{C} w_c \mathcal{N}(o|\mathbf{m}_c + \mathbf{D}_c\mathbf{z}_c, \mathbf{\Sigma}_c) \qquad (1)$$

The mean vectors $\mathbf{m}_c + \mathbf{D}_c\mathbf{z}_c$ of the GMM are obtained via the *Maximum a Posteriori* (MAP) criterion, or more precisely the relevance MAP, in which $\mathbf{D}_c^\intercal\mathbf{D}_c = \tau^{-1}\mathbf{\Sigma}$. Here, τ is the relevance factor, the value of which is set to 2 in our implementation. Only the mean vector are adapted, while the remaining parameters are taken to be the same as that of the UBM $\lambda_{ubm} = \{w_c, \mathbf{m}_c, \mathbf{\Sigma}_c; c = 1, 2, \ldots, C\}$ consisting of the weight w_c, mean vectors $\mathbf{m}_c$, and covariance matrices $\mathbf{\Sigma}_c$. The GMM is then used to adapt the five states of a left-to-right HMM by using the MAP criteria. Let s be the state index, the state emission probability is obtained via:

$$p(o|s) = \sum_{c=1}^{C} w_c \mathcal{N}(o|\mathbf{m}_c + \mathbf{D}_c\mathbf{z}_c + \mathbf{D}_c\mathbf{z}_{c,s}, \mathbf{\Sigma}_c) \qquad (2)$$

A detailed description of the HiLAM is given in [20, 21]. The choice of a system that does not include any channel compensation technology [1] is motivated by our intention to demonstrate the magnitude of the VHF channel degradation.

4.2. Experimental protocol

The aim of this work is to study the effect of the VHF channel on the performance of the speaker verification system. Due to bandwidth limitation of the VHF channel, under 4kHz, we created a down-sampled version of the *RSR2015* reducing the bandwidth by half for comparison. Therefore, experiments are reported, when possible, on the original 16kHz database, on its down-sample version at 8kHz and on the corresponding data transmitted through the VHF channel.

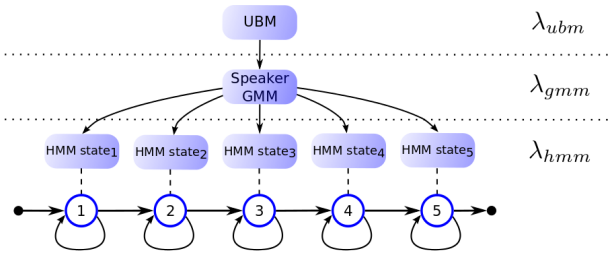

Figure 3: Three layer acoustic architecture of the Hierarchical multi-Layer Acoustic Model (HiLAM).

The 300 speakers are divided into three non-overlapping groups labeled *background* (50 male / 47 female), *development* (50 male / 47 female) and *evaluation* (53 male / 49 female). We recommend to use the *background* for estimation of the background parameters such as training of the UBM. The *development* set can be used to estimate the decision threshold and possible calibration parameters while the *evaluation* set is used for validation.

As described in Section 2, all speaker recorded their 9 sessions by using 3 different devices in sequence. For each speaker, sessions $\{1, 4, 7\}$ recorded on the same device are used for enrollment. The 6 remaining sessions are used for testing. In order to keep the enrollment session below 10s, only 3 occurrences of a pass-phrase are used to adapt both the middle-layer GMM and the pass-phrase HMM. The 30 user- and pass-phrase-dependent models trained for each speaker are tested against all utterances from the 6 test sessions of all other speakers from the same gender and data-set (i.e., *development* or *evaluation*). Note that for consistency, this experimental protocol is identical to the one used in [11] where total numbers of models and trials per condition can be find.

5. Benchmark and analysis

The first set of experiments is performed to compare the degradation caused by the down-sampling of the signal from 16kHz to 8kHz and by the VHF channel. Table 1 summarizes the performance of the HiLAM system for male and female speakers on both the *development* and *evaluation* sets. From these results, it appears that down-sampling the signal from 16kHz to 8kHz (where the bandwidth is reduced by half from 8kHz to

4kHz) helps for the case of rejecting a user, target or impostor, pronouncing a pass-phrase different from the one used in test. This is the case for all conditions except the target female from the *evaluation* set where the difference between equal error rates in 16kHz and 8kHz is only 0.02% absolute and is not significant. These results are counter intuitive as we expect to get more information from the signal with wider bandwidth. One possible reason is that for this specific task of pass-phrase recognition, down-sampling the signal removes part of the speech information but also a larger part of the characteristic information of the speaker that might *hide* the pass-phrase for the case of our system which models both speaker and pass-phrase within a single HMM [22, 23].

For the case of an impostor pronouncing the correct pass-phrase (the one chosen by the target speaker during the enrollment), down-sampling of the speech signal degrades the performance of the system in all conditions except male *evaluation*. It can also be noticed that the effect of down-sampling is more for female speakers as formant frequencies are higher. The EER increases by 61.8% on the *development* set and 120% on the *evaluation* set while it only increases by 12.5% on the male *development* set.

As expected, transmission through the VHF radio channel generates severe degradations in terms of accuracy. EERs obtain after transmission through VHF is at least 5 times higher than when using the equivalent 8kHz data and 6.5 times higher than the equivalent measured on the original 16kHz data. This experiment demonstrates that the major degradation of performance is due to the alteration of the spectrum by the VHF channel.

An important result of this experiment comes from the comparison of two test conditions, namely, the conditions in which the system has to reject a target speaker pronouncing a wrong pass-phrase (*TAR-wrong*) or an impostor pronouncing the correct one (*IMP-correct*). *TAR-wrong* consists of a pass-phrase recognition task while *IMP-correct* is a speaker verification task with a perfect lexical match. For both 16kHz and 8kHz signal, in all 4 conditions (male, female, *development*, *evaluation*), the EER against *TAR-wrong* is at least 30% lower than the one obtained against *IMP-correct*, i.e., it is easier to reject a wrong pass-phrase than an impostor. However, in the case where the speech signal has been transmitted over VHF, the trend changes and it is easier for the system to reject an impostor than a wrong pass-phrase. Again, this result is counter-intuitive and requires more analysis as the VHF channel was expected to alter more the highest frequencies and thus affect more the speaker verification task.

The second set of experiments is performed to measure the degradation due to cross-channel testing. Tables 2 and 3 summarize the performance of the HiLAM system when enrollment is done on data after down-sampling to 8kHz or transmission over VHF and tests are performed on the same two conditions on the male and female *development* part of the *RSR2015* database. As expected, cross-channel testing increases both EER and minDCF in a dramatic manner. The condition where impostors pronounce a wrong pass-phrase is especially affected as EER increases by at least 152% for the male case compare to the case where both enrollment and test data have been transmitted over VHF. The relation between pass-phrase and speaker specific information seems to be extremely complex. Indeed, in the case of cross-channel testing, the EER obtained for impostors pronouncing a correct pass-phrase is higher than the one obtained against target speaker pronouncing a wrong pass-phrase, as it is the case for a clean channel (16kHz and 8kHz) but not for the case where both enrollment and tests are performed over VHF channel.

Table 2: Performance of *HiLAM* system on the male *development* set of Part I of the *RSR2015* in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of non-target access and cross channel conditions.

		Channel		
			Test	
User	Text	Enrollment	8kHz	VHF
Target	Wrong	8kHz	0.75 / 0.27	20.03 / 7.47
		VHF	20.03/ 6.77	16.10 / 8.23
Impostor	Correct	8kHz	1.71 / 0.77	22.55 / 8.20
		VHF	23.59/ 8.23	10.95 / 4.95
Impostor	Wrong	8kHz	0.10 / 0.03	15.02 / 6.10
		VHF	16.21/ 5.30	5.79 / 2.64

Table 3: Performance of *HiLAM* system on the female *development* set of Part I of the *RSR2015* in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of non-target access and cross channel conditions.

		Channel		
			Test	
User	Text	Enrollment	8kHz	VHF
Target	Wrong	8kHz	0.44 / 0.24	19.11 / 6.97
		VHF	18.40/ 6.62	12.46 / 7.24
Impostor	Correct	8kHz	1.53 / 0.76	22.11 / 7.88
		VHF	21.42/ 7.60	9.01 / 4.26
Impostor	Wrong	8kHz	0.02 / 0.01	14.73 / 5.78
		VHF	15.11/ 5.07	4.40 / 2.06

6. Discussion

We have presented an extension of the *RSR2015* database for text-dependent speaker verification over VHF channel and some initial results and analysis. We described the platform used for transmit-and-record of the original data over the VHF channel in order to help future implementations of such process. We have compared performance of a text-dependent speaker verification engine on the original 16kHz data, on the same data down-sampled to 8kHz, after transmission over the VHF channel and in cross-conditions. As expected, the VHF channel strongly affects the accuracy of the speaker verification engine and performance are even worse in the case of cross-channel testing. Our experiments shows that bandwidth limitation and VHF channel alter the speaker verification and pass-phrase verification in different manners that are not yet fully understood and will require more analysis in the future.

7. Acknowledgement

We would like to thank Shelley L.Y. Loh for her efforts in developing the VHF transmit-and-record platform.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] M. Hébert, *Text-dependent speaker recognition*. Springer-Verlag, Heidelberg, 2008.

[3] D. A. Reynolds, "An overview of automatic speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2002, pp. 4072–4075.

[4] NIST, "The nist year 2012 speaker recognition evaluation plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf, 2012.

[5] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.-M. Bousquet, E. Khoury, P. L. S. Martinez, J. M. K. Kua, C. H. You, H. Sun, A. Larcher, P. Rajan, V. Hautamaki, C. Hanilci, B. Braithwaite, R. Gonzales-Hautamaki, S. O. Sadjadi, G. Liu, H. Boril, N. Shokouhi, D. Matrouf, L. E. Shafey, P. Mowlaee, J. Epps, T. Thiruvaran, D. A. van Leeuwen, B. Ma, H. Li, J. H. L. Hansen, J.-F. Bonastre, S. Marcel, J. S. Mason, and E. Ambikairajah, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.

[6] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey Speaker and Language Recognition Workshop*, 2012, pp. 1–7.

[7] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Ma, S. Cumani, O. Glembeck, H. Hermansky, S. Mallidi, N. Mesgarani *et al.*, "Developing a Speaker Identification System for the Darpa RATS Project," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013.

[8] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving Speaker Identification Robustness to Highly Channel-Degraded Speech Through Multiple System Fusion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2013.

[9] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," *Signal Processing Letters*, vol. 20, pp. 197–200, 2012.

[10] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM Speech Activity Detection System for the DARPA RATS Program," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 3497–3501.

[11] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent Speaker Verification: Classifiers, Databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[12] R. Cole, M. Noel, and V. Noel, "The CSLU speaker recognition corpus," in *Proceedings International Conference on Spoken Language Processing, ICSLP*, 1998, pp. 3167–3170.

[13] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree *et al.*, "The BANCA database and evaluation protocol," *Lecture Notes in Computer Science (LNCS)*, vol. 2688, pp. 625–638, 2003.

[14] B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, and D. V. Rotz, "MyIdea–Multimodal biometrics database, description of acquisition protocols," *Biometrics on the Internet*, vol. 275, pp. 59–62, 2005.

[15] R. H. Woo, A. Park, and T. J. Hazen, "The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments," in *Odyssey Speaker and Language Recognition Workshop*, 2006.

[16] H. Meng, P. Ching, T. Lee, M. W. Mak, B. Mak, Y. Moon, X. Siu, M.-H. Tang, X. Tang, H. P. Hui, A. Lee *et al.*, "The multibiometric, multi-device and multilingual (m3) corpus," in *International Workshop on Multimodal User Authentication*, 2006, pp. 1–8.

[17] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012, pp. 1580–1583.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus linguistic data consortium," *Philadelphia, PA*, vol. 1, 1993.

[19] L. Mercz, *Marine VHF Radio Handbook*. Mercator, 2010.

[20] A. Larcher, J.-F. Bonastre, and J. S. Mason, "Reinforced temporal structure of acoustic models for speaker recognition," *Digital Signal Processing*, vol. 23, no. 6, pp. 1910–1917, December 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200413001504

[21] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Imposture classification for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2014, pp. 739–743.

[22] L. Besacier, J.-F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2-3, pp. 89–106, 2000.

[23] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Care, "Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification," in *IEEE Signal Processing Letters*, vol. 19, no. 12, 2012.