Speaker age estimation for elderly speech recognition in European Portuguese

Thomas Pellegrini[1]*, Vahid Hedayati*[2]*, Isabel Trancoso*[2,3]
Annika Hämäläinen[4]*, Miguel Sales Dias*[4]

[1]Université de Toulouse; UPS; IRIT; Toulouse, France
[2]INESC-ID, Lisbon, Portugal
[3]Instituto Superior Técnico, Lisbon, Portugal
[4] Microsoft Language Development Center, Lisbon, Portugal /
ISCTE - University Institute of Lisbon (ISCTE-IUL), Lisbon, Portugal
pellegri@irit.fr, vahidhdyt@gmail.com, isabel.trancoso@inesc-id.pt,
t-anhama@microsoft.com, miguel.dias@microsoft.com

Abstract

Phone-like acoustic models (AMs) used in large-vocabulary automatic speech recognition (ASR) systems are usually trained with speech collected from young adult speakers. Using such models, ASR performance may decrease by about 10% absolute when transcribing elderly speech. Ageing is known to alter speech production in ways that require ASR systems to be adapted, in particular at the level of acoustic modeling. In this study, we investigated automatic age estimation in order to select age-specific adapted AMs. A large corpus of read speech from European Portuguese speakers aged 60 or over was used. Age estimation (AE) based on i-vectors and support vector regression achieved mean error rates of about 4.2 and 4.5 years for males and females, respectively. Compared with a baseline ASR system with AMs trained using young adult speech and a WER of 13.9%, the selection of five-year-range adapted AMs, based on the estimated age of the speakers, led to a decrease in WER of about 9.3% relative (1.3% absolute). Comparable gains in ASR performance were observed when considering two larger age ranges (60-75 and 76-90) instead of six five-year ranges, suggesting that it would be sufficient to use the two large ranges only.

Index Terms: automatic speech recognition, elderly speech, automatic age estimation, i-vector extraction

1. Introduction

In the context of a Portuguese national project called "AVoz", we studied European Portuguese (EP) elderly speech with the objective of improving speech recognition for elderly speakers. There is no standard age boundary to define the elderly. However, to give an idea, speakers aged above 75 are often called elderly speakers in the literature. Independent of the language in question, speech recognizers' performance is significantly worse in the case of elderly speech than in the case of young adult speech ([1, 2, 3]). There are several reasons for this. First, some parameters of the speech signal (e.g. speech rate, F0, jitter, shimmer) change with age ([4, 5, 6]), while the acoustic models of speech recognizers are typically trained using speech from younger adults, with elderly speakers not appearing at all or being under-represented in the training data. Second, the elderly usually interact with computers using everyday language and their own commands, even when a specific syntax is required ([7]). Improvements in ASR performance can be achieved by using acoustic models (AMs) specifically adapted to the elderly ([3, 8]). In order to automize the selection of age-specific adapted AMs, one could estimate the age of the speakers automatically.

A number of studies have explored automatic age estimation. In an early study, Minematsu et al. ([9]) estimated speakers' age only using acoustic (i.e., no linguistic) information. They used Gaussian Mixture Models (GMMs) to distinguish between two groups of speakers defined using the results of previous listening tests: speakers whose speech sounded very old to the judges ("subjective elderly") and a control group with the rest of the speakers in their databases ("non-subjective elderly"). A correct automatic identification rate of 91% was achieved, and the rate further increased to 95% when using additional prosodic features. More recent studies have continued to use techniques derived from the speaker recognition field, such as GMM supervectors ([10, 11, 12]), and, more recently, i-vectors, based on the so-called total variability model ([13]). These techniques involve estimating vectors that somehow characterize the speaker's voice. Thus, age is a factor that may also be represented in the vectors. Channel compensation techniques may be applied to the vectors to focus on the speaker characteristics only. In these studies, after gathering the supervectors or the i-vectors, Support Vector Machines (SVM) for classification or regression are used to estimate either the age range or the specific age of the speaker. i-vectors have the advantage of producing low-dimension vectors, typically between 200 and 400. In [13], their use also resulted in the best AE performance, as compared with the GMM-supervector technique that had a mean absolute error (MAE) of 7.6 years.

In this study, after giving a bird's-eye view of AE techniques in Section 2, we report experiments on a subset of a large corpus of read elderly speech in European Portuguese. The corpus is described in Section 3. We used AE to select age-specific adapted AMs and report the results of our ASR experiments in Section 5. To the best of our knowledge, apart from another study of ours [14], previous studies on AE do not report ASR experiments exploiting the results of the AE.

2. Automatic age estimation

Until recently, the GMM supervector paradigm was the state-of-the-art technique in the field of speaker recognition. In a nutshell, this approach consists of training a Universal Background

Table 1: *Main statistics of the speech material*

Set	Gender	# Spk	Duration	# Word	
				Types	Tokens
Training	female	528	6h30	4.4k	32.3k
	male	179	2h17	2.5k	11.6k
Test	female	225	2h43	2.7k	12.2k
	male	75	1h01	1.4k	5.3k

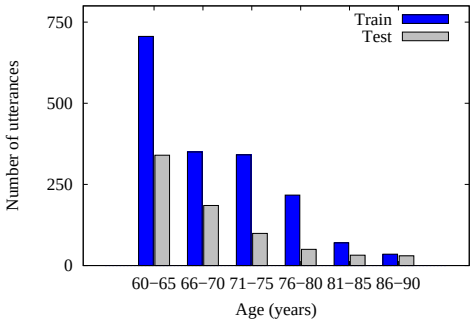

Figure 1: Histograms of the number of utterances in the training and test sets containing male speakers.

Model (UBM) with a large set of speakers and then performing UBM adaptation (usually using Maximum A Posteriori (MAP) adaptation) to gather speaker-dependent high-dimensional supervectors, composed of the concatenated means of the adapted GMMs.

Significant improvements have been obtained using a new technique, referred to as the total variability approach, in which *i-vectors* are extracted ([15]). Instead of using GMM supervectors as such, with standard dimensions greater than 10k, this new approach proposes to represent channel variability and speaker characteristics simultaneously with a low dimensional sub-space called the total variability sub-space. Speech utterances can be projected onto this sub-space to be represented by *i-vectors*, which have a low dimension that is typically between 200 and 400.

Both the GMM-supervector and the i-vector approaches have already been used for AE in various studies, such as [10, 11, 13]. In [13], the authors used the GMM-supervector approach. Five age classes were used to estimate the age of children between five and ten years of age. The authors used GMM-supervectors and an SVM classifier. Overall precision and recall of 83% and 60% were achieved. The authors also reported slightly worse results when using Support Vector Regression (SVR). However, SVR produced more balanced results among the different age ranges. This approach is similar to the one used in [10], except that the authors compare the effect of both the GMM-supervector and the i-vector approaches on AE performance, on large sets of telephone speech data. The i-vector approach outperformed the GMM-supervector approach, with MAE rates of 7.6, and 7.9 years, respectively.

Seeing that the total variability approach is state-of-the-art both in speaker recognition and in AE, we adopted the i-vector/SVR approach also for our study.

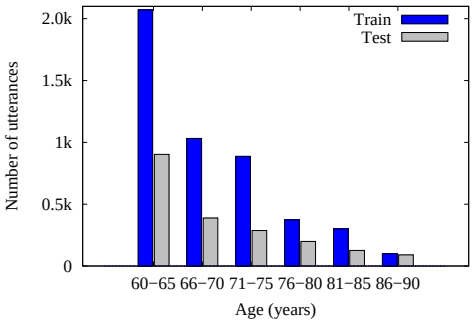

Figure 2: Histograms of the number of utterances in the training and test sets containing female speakers.

3. Speech material

As in our previous work, we used the EASR Corpus of European Portuguese Elderly Speech ([16]). The corpus contains about 190 hours of read speech, including silences. A total of about 1000 speakers aged 60 or over read out 160 prompts representing 14 different prompt types ranging from isolated digits to phonetically rich sentences. The exact age of the speakers is not known; the age of the speakers is reported using five-year ranges: 60-65, 66-70 and so on. 72% of the speakers in the corpus are female.

The number of speakers and the duration of the data used in this study are presented in Table 1. For this work, we only used 10 utterances per speaker, corresponding to about one minute of speech per speaker. We limited the amount of speech per speaker to have an experimental setup similar to that used in NIST speaker identification evaluations, in which the amount of speech is limited to 20-160 seconds per speaker ([17]). The training and test sets contained about 70% and 30% of the utterances in the whole corpus, corresponding to almost 9h and 3h45 of speech, respectively. None of the speakers in the training set appear in the test set.

Figures 1 and 2 show histograms of the number of utterances used for training and testing in the case of male and female speakers, respectively. Only six five-year ranges were considered: 60-65, 66-70, 71-75, 76-80, 81-85, and 86-90. Older speakers were not considered due to lack of data. The age range and gender distributions in the full corpus were respected when creating the training and test sets. As can be seen in the figures, speakers in the 60-65 age range are the most numerous. In the training sets, there are 706 and 2,070 utterances from male and female speakers in that age range, respectively, as compared with 35 and 99 utterances from the oldest male and female speakers (in the 86-90 age range).

4. Age estimation results

Gender-dependent UBMs of 1024 Gaussian mixtures were trained on the male and female training sets. The acoustic features consisted of 13 Mel-Frequency Cepstral Coefficients (MFCCs), including energy, with their first order derivatives, resulting in 26-d feature vectors extracted every 10 ms with 20 ms Hamming window frames. Energy-based speech activity detection was used, followed by mean-variance feature normalization. 200-d i-vectors were extracted for each utterance, both for the training and the test sets. The ALIZE toolbox was used to

Table 2: *Age estimation results (Mean Absolute Error rates in years) on the test sets.*

	6 age ranges balanced		2 age ranges balanced	
	no	yes	no	yes
Males	5.45	9.63	4.18	8.60
Females	5.68	8.32	4.53	7.06

perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set for the 86-90 age range (with the smallest amount of data): 35 vectors for males, and 100 for females. In other words, in this experiment, the AE training corpus is balanced in terms of age distribution. This change increased the global MAE values to 9.63 and 8.32 years for males and females, respectively. Nevertheless, as can be seen in Figure 3, which shows the MAE histograms for five-year age ranges using the original ("unbalanced") and the "balanced" set of i-vectors to train the SVR, the error rates are centered around the age of 75 years when using the balanced training set. For male speakers, the MAEs with the original unbalanced training set were 5.1 and -19.5 years for the 60-65 and 86-90 ranges, respectively. For the same test set and the same age ranges but using the down-sampled balanced set to learn the total variability matrix, the MAEs were 10.7 and -11.9 years. Fewer errors are made for the speakers of the test set aged between 71 and 75. When only using two large age ranges, the global MAE decreased to 8.60 and 7.06, as was already observed in the non-balanced configuration. Since the oldest speakers are less represented in the test set, just like in the original training set, the impact of balancing the training set increased the estimation errors made on the youngest speakers, therefore increasing the global error rates. Finally, slightly better AE figures were obtained for females, which might be consistent with the better ageing compensation in ageing female speakers as reported in [20]. This could be linked to more consistent physiological traits of ageing in females than in males, although this remains to be investigated. In the remainder of this study, we used the age estimation module that resulted in the smallest error rates, *i.e.* the one trained with the original unbalanced data. In the next section, ASR performance is compared when using two and six age ranges.

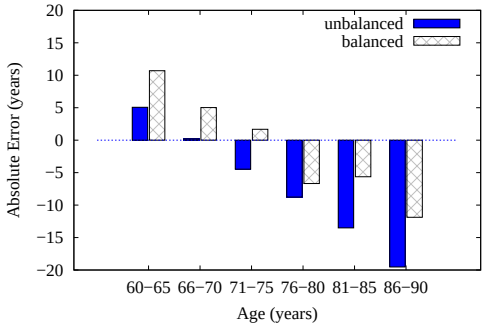

Figure 3: Histograms of the age estimation error rates obtained with six age ranges on the test set containing male speakers, in terms of Mean Absolute Error rates (in years). Blue solid fill: original age-unbalanced training set, light Gray fill: age-balanced training set.

5. Application to ASR

The speech recognizer used for the study, Audimus, is a hybrid speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs) ([21]). The MLPs perform a phoneme classification by estimating the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated with a single state of context-independent phoneme HMMs. More specifically, the system combines three MLP outputs trained with Perceptual Linear Prediction (PLP) features (13 static + first derivative), log-RelAtive SpecTrAl (RASTA) features (13 static + first derivative) and Modulation SpectroGram (MSG) features (28 static) ([22]). The MLP parameters were initially trained with 46 hours of manually transcribed television news broadcasts and then with 1000 hours of automatically transcribed television news broadcasts and selected according to a confidence measure threshold (non-supervised training). These data feature speech from young and middle-aged adult speakers mainly from the Lisbon area. The MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three-state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units. The Word Error Rate (WER) of our ASR system is, on average, under 20% for broadcast news (BN) speech ([23]). For this study, we used a 3-gram language model (LM) with Kneser-Ney modified smoothing learned on the training set of the corpus. We also report results with a 2-gram model. The multiple-pronunciation EP lexicon used in this study includes about 114k entries. The out-of-vocabulary rate for the EASR transcripts is below 1%. The baseline AMs were trained using BN data.

In a previous study, we explored AM adaptation with the EASR Corpus ([8]). The baseline MLPs were re-trained with age-specific data from the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age ranges from 60-65 to 86-90. Six hours of speech extracted from the training set were used to adapt the baseline MLP to

Table 3: *ASR performance in terms of WER (%). Baseline: AMs trained with young adult speech, oracle: adapted AMs and ground-truth age (6 or 2 age ranges), estimated: adapted AMs and estimated age (6 or 2 age ranges).*

LM	baseline	oracle		estimated	
		6 AMs	2 AMs	6 AMs	2 AMs
2g	21.3	18.1	18.7	20.6	19.1
3g	13.9	11.5	12.6	13.4	12.1

each of the five-year age ranges, except for the last one, 86-90, for which only two hours of data were available. Using adaptation data from the oldest speakers gave better results on the test sets of the oldest speakers. For instance, AM-60-65 and AM-81-85 respectively showed 13.7% and 22.0% relative improvements over the baseline for the 81-85 test set (5.6% and 9.0% absolute, respectively). However, in that study, we did not use AE to automatically select the age-specific adapted models.

In this study, we wanted to measure the performance decrease when estimating speaker age automatically, as compared with an *oracle* system, in which the real age of the speaker is known *a priori*. We also wanted to find out whether or not it is worth using six five-year-age ranges or if two larger age ranges would be sufficient. Table 3 shows the WER values for all the setups for the male speakers. The performance obtained in the case of the female speakers is not reported, as the WERs are similar, although slightly lower.

Baseline WERs of 21.3% and 13.9% were achieved on our test set with a 2-gram and a 3-gram LM, respectively. When using the AMs adapted with speech data from the age range corresponding to the real age of each test speaker, an experimental setup that corresponds to the *oracle* columns in Table 3, the WER decreased to 18.1% and 11.5% (about 17.0% relative) with six age ranges, and to 18.7% and 12.6% with two age ranges.

When automatically estimating speaker age to select the adapted AMs, performance decreased by about 2% absolute as compared with the oracle setup. A decrease could indeed be expected because of age estimation errors. One result contradicts this hypothesis, though: a lower 12.1% global WER was obtained with the 3-gram LM and the two age ranges than the 12.6% obtained with the oracle setup. Some speakers aged 76-90 were incorrectly classified as 60-75-year-old speakers. Therefore, the AMs adapted for the latter age range may model the characteristics of their speech better in the sense that their real age does not really correspond to "the age of their speech". However, a similar result was not obtained with the 2-gram LM. Hence, a more probable reason for the result is an effect of the small size of the 75-90 test set that appeared when using a larger order LM (3-gram LM), which requiers more data to be correctly trained.

Interestingly, the performance differences are small when using six or just two sets of adapted AMs. Thus, it may be easier to only use two age ranges, 60-75 and 75-90, when adapting AMs for the elderly.

6. Conclusions

In this study, we investigated automatic age estimation as a preprocessing step to selecting age-specific AMs for elderly speech recognition. Our objective was to verify that AE based on i-vectors is efficient enough for a model selection frontend in ASR. A positive result, which we obtained, suggests that it might not be necessary for elderly users to spend time adapting ASR systems to their voices. We used a large corpus of read speech from European Portuguese speakers aged 60 or over. The exact age of the speakers in the corpus is not known; the age of the speakers is reported using five-year age ranges. We performed AE using i-vectors and support vector regression. When using six five-year age ranges from 60-65 to 86-90, we obtained mean error rates of about 5.4 and 5.7 years for male and female speakers, respectively. When only using two larger age ranges, 60-75 and 75-90, the error rates decreased to 4.2 and 4.5. The selection of five-year-range adapted AMs, based on the automatically estimated age ranges of the test speakers, led to a decrease in WER of 9.3% relative (1.3% absolute) over the WER of 13.9% obtained using a baseline ASR system without AM adaptation to elderly speech. Only small differences were observed when only using the two larger age ranges. Thus, it only seems necessary to use two sets of adapted AMs to recognize speakers aged 60 and over.

In future work, we will compare these results with speaker adaptation in the context of a hybrid HMM/MLP system. Furthermore, in collaboration with the Microsoft Language Development Center in Lisbon, we are currently carrying out AE experiments with speakers representing a wide range of ages from three years old until old age [14].

7. Acknowledgments

This work was partially supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under projects PTDC/EEA-PLP/121111/2010, PTDC/EIA-CCO/122542/2010, and PEst-OE/EEI/LA0021/2011, and also by the QREN 5329 Fala Global project, which is co-funded by Microsoft and the European Structural Funds for Portugal (FEDER) through POR Lisboa (Regional Operational Programme of Lisbon), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry.

8. References

[1] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, Atlanta, 1996, pp. 349–352.

[2] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Acoustic models of the elderly for large-vocabulary continuous speech recognition," *Electronics and Communications in Japan*, vol. 87:7, pp. 49–57, 2004.

[3] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proc. Interspeech*, Brisbane, 2008, pp. 2550–2553.

[4] S. Xue and G. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study," *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 689–701, 2003.

[5] C. Reynolds and J. Czaja, S.and Sharit, "Age and perceptions of usability on telephone menu systems," in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 2, 2002, pp. 175–179.

[6] T. Pellegrini, A. Hämäläinen, P. Boula de Mareüil, M. Tjalve, I. Trancoso, S. Candeias, M. Dias, and B. Braga, "A corpus-based study between elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance," in *To appear Proc. Interspeech*, Lyon, 2013.

[7] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval," in *Proc. ICASSP*, Phoenix, 1999, pp. 145–148.

[8] T. Pellegrini, I. Trancoso, A. Hämäläinen, A. Calado, M. Dias, and D. Braga, "Impact of age in ASR for the elderly: preliminary experiments in European Portuguese," in *Proc. IberSPEECH*, Madrid, 2012.

[9] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," in *Proc. ICASSP*, Orlando, 2002, pp. 137–140.

[10] T. Bocklet, A. Maier, and E. Nth, "Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horak, I. Kopecek, and K. Pala, Eds. Springer Berlin Heidelberg, 2008, vol. 5246, pp. 253–260.

[11] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.

[12] D. Mahmoodi, A. Soleimani, H. Marvi, F. Razzazi, M. Taghizadeh, and M. Mahmoodi, "Age estimation based on speech features and support vector machine," in *Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd*, Colchester, 2011, pp. 60–64.

[13] M. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," in *Proc. Interspeech*, 2012.

[14] A. Hämäläinen, H. Meinedo, M. Tjalve, T. Pellegrini, I. Trancoso, and M. Sales Dias, "Improving Speech Recognition through Automatic Selection of Age Group Specific Acoustic Models," in *to appear in Proc. PROPOR*, São Carlos, 2014.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[16] A. Hämäläinen, J. Rodrigues, M. Sales Dias, A. Kolesinski, T. Fegyó, G. Németh, P. Csobánka, K. Lan Hing Ting, and D. Hewson, "The EASR Corpora of European Portuguese, French, Hungarian and Polish Elderly Speech," in *Proc. LREC*, Reykjavik, 2014.

[17] "The NIST Year 2012 Speaker Recognition Evaluation Plan," 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm

[18] J. F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. ICASSP*, 2005, pp. 737–740.

[19] M. e. a. Hall, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11:1, 2009.

[20] F. Kelly, N. Brümmer, and N. Harte, "Eigenageing compensation for speaker verification," in *Proc. Interspeech*, Lyon, 2013, pp. 1624–1628.

[21] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. ICASSP 2008*, Las Vegas, USA, 2008.

[22] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a broadcast news speech recognition system for the european portuguese language," in *proceedings of PROPOR*, Faro, 2003, pp. 9–17.

[23] H. Meinedo, A. Abad, T. Pellegrini, J. Neto, and I. Trancoso, "The L2F Broadcast News Speech Recognition System," in *Proc. Fala*, Vigo, 2010, pp. 93–96.