A STUDY OF AUTOMATIC PHONETIC SEGMENTATION FOR FORENSIC VOICE COMPARISON

Chee Cheun Huang[1,2]*and Julien Epps*[1,2]
[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Sydney, NSW 1430, Australia
chee.huang@student.unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

Forensic voice comparison (FVC) systems have often involved manual annotation of usable phonetic units, requiring substantial human labor. Recent research has shown the efficacy of automatic methods in FVC, and this paper investigates automatic phonetic segmentation in FVC systems. Nasals and vowels were found to contribute the most in terms of improvements in both the validity and reliability of the system. Results show that as a function of the duration of the recognized tokens there is a trade-off in which an improvement in validity corresponds to a degradation in reliability and vice versa. An implication is that minimizing the error of automatically estimated monophone boundaries may not necessarily result in the best system validity or reliability. A substantial improvement in log-likelihood-ratio cost (validity) of 17.02% and in 95% credible interval (reliability) of 5.97% over the baseline system was possible by fusing baseline scores with those from nasal and vowel segments.

Index Terms— Forensic Voice Comparison, log-likelihood ratio cost, 95% credible interval, validity, reliability

1. INTRODUCTION

FVC systems have often employed the acoustic-phonetic approach, which is characterized by trained phoneticians identifying and marking sufficient usable tokens of phonetic units in both the suspect and offender recordings, which are then subjected to statistical analysis. Forensic science is undergoing a shift towards a new paradigm, characterized by a quantitative data-based implementation of the likelihood ratio framework [1], together with evaluation of the validity and reliability of results. In this new paradigm, there is an increasing research interest in automation of FVC systems that could result in reduction in human labor cost and hence increase in system efficiency. In particular, recent FVC research [2] has shown the value of combining automatic systems with acoustic-phonetic systems with improvements in both validity and reliability (i.e. the confidence the trier of fact can have in the likelihood ratio presented).

Studies of automatic *speaker recognition* in the past have shown that nasals and vowels are effective in system performance. In particular, liquids and glides, vowels, and nasal vowels and consonants were found to contain more speaker-specific information and provide better speaker identification results than phonetically balanced test utterances [3]. Similar results were observed from experiments based on Vector Quantization (VQ) methods [4] and autoregressive vector models [5] respectively. Frames from syllables whose corresponding word-level speech recognition transcript included the letter 'n' performed better relative to a baseline system which used all frames [6]. These studies were performed on the basis of ranking system performance in terms of accuracy metric without considering precision. Further, there was no systematic analysis of the effect of errors between the temporal locations of the true and automatically detected phonetic tokens on the performance of automatic systems.

In the present paper, the effectiveness of HMM-based phone recognition for forensic voice comparison is evaluated in terms of both validity (accuracy) and reliability (precision). The main motivation for this study is to address the following questions: (1) In forensic acoustics, manual segmentation has traditionally been considered as the most accurate method of detecting usable phonetic tokens. This level of accuracy cannot be reproduced using current automatic methods. However as an initial step towards reducing human labor, what is the best automatic method that could be used in forensic work ? (2) In a typical forensic offender recording, the duration and availability of phonemes within the recording is often limited. What are the phones that we should consider in order to achieve the best possible FVC system performance under this condition ? (3) How sensitive is a FVC system to errors in the endpoints of automatically detected phonetic tokens ?

It must be emphasized that this paper does not aim to replicate past phone segmentation studies done within the context of automatic speaker verification; the variation in speaker discrimination offered by different classes of phonemes is already established [7, 8, 9]. The context of this work relates to automatic speech segmentation in a scenario closer to realistic forensic casework conditions. Since databases replicating such conditions are still under development, recordings used in the present study were carefully chosen to adhere to the characteristics (particularly the gender, language and recording duration) of the suspect and offender recordings. Further, Detect Error Trade-off (DET) curve and Equal Error Rate (EER), which originate from the classical false acceptance/false rejections (FA/FR) performance metrics for speaker recognition system, are not appropriate within the likelihood ratio framework, and are replaced by the more appropriate the log-likelihood-ratio cost (C_{llr}) and 95% Credible Interval (95% CI), see [1] and Appendix B of [10].

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

2. METHODOLOGY

2.1. Automatic Phonetic Segmentation

A Hidden Markov Model Toolkit (HTK) based phone recognizer was constructed. The TIMIT database of read speech was used to train and evaluate the phone recognizer.

The phones selected for modeling were the same as those listed in [11] with some minor modifications. Briefly, there were a total of 41 phones and 3 TIMIT symbols (ie 'cl' (unvoiced closure), 'vcl' (voiced closure) and 'sil' (silence)) selected in our study. Four phones (/ʔ/, /ə^{h}/, /ɨ/ and /y/) were not included, and 17 allophones were combined with their corresponding TIMIT phoneme symbols. Each of the 41 phones and 3 TIMIT symbols was then represented by an HMM, implemented using HTK, containing 3 states with 12 mixtures per state. The HMMs were trained as right-context-dependent biphone models. Finally, recognition was carried out via Viterbi search with bigram language models, producing monophone phonetic transcriptions as output. The phone recognizer was trained on TIMIT, attaining a 44-class phoneme accuracy of 63% on the TIMIT test database.

2.2. Forensic Voice Comparison System Configuration

The automatic FVC system was the same as that employed in [12]. Briefly, a 512-mixture component Gaussian Mixture Model – Universal Background Model (GMM-UBM) was used to model 32-dimensional MFCCs (16 static coefficients and 16 delta coefficients) extracted from 20ms frames overlapped by 10ms. It is common in speaker recognition to train a Universal Background Model (UBM) that models the acoustic space of all speakers using a large database. This is typically followed by maximum a posteriori (MAP) adaptation to derive the speaker model from the trained UBM parameters [17]. The UBM can be considered as a model covering the entire broad acoustic space for all speaker independent acoustic classes, while MAP adaptation tunes these acoustic classes based on the speaker dependent training data. For the unseen acoustic classes in the speaker dependent training speech, the adapted mixture parameters are copied directly from the UBM. This will result in an almost zero log-likelihood ratio (i.e. neither support nor against the hypothesized speaker) in the recognition phase for those acoustic classes that are unseen in the training data [17]. While GMM-UBM has been widely used in automatic *speaker recognition*, studies in the area of forensic acoustics have also recently adopted this technique [18].

Feature normalization was performed via cumulative distribution mapping and no channel or session compensation technique was applied. Prior to phone recognition, an energy-based Voice Activity Detector (VAD) was applied to remove portions of the recording which the speaker was not talking [13].

2.3. Background, development, and test databases

The UBM was trained from the 750 longest recordings of US English male speakers from the NIST SRE 2006 8conv database The total number of speakers was 101, with 4 to 8 recordings per speaker. Development data used to train calibration and fusion weights consisted of two non-contemporaneous recordings (with speech-active duration ranging from 84s to 131s with a median of 110s) of each of 32 male speakers of US English from the NIST SRE 2008 8conv database. Test data consisted of four non-contemporaneous recordings (with speech-active duration ranging from 84s to 159s with a median of 109s) of 100 male speakers of US English from NIST 2008 8conv. The database is described in more detail in [12], however it should be emphasized that recordings for the background, development and test databases were chosen to approximate the characteristics of forensic casework, since there is currently no large database of this kind available. Weights for logistic regression calibration and fusion [14, 15, 16] were calculated using the scores derived from the development data, and these weights were then used to calibrate and fuse scores from the test data. The pooled procedure for calculating the weights was adopted (see [12]).

2.4. Calculation of validity and reliability

The validity and reliability (accuracy and precision) of the FVC systems were evaluated using the log-10-likelihood-ratio cost (C_{llr}) and a parametric estimate of the log-10-likelihood-ratio 95% credible interval (95% CI) respectively, see [1].

3. RESULTS

3.1. Phonetic segmentation results

A scatter plot of reliability versus validity for each system is displayed in Figure 1. Three infrequently occurring phones: /p/, /ʧ/ and /θ/ and three TIMIT symbols: 'cl' (unvoiced closure), 'vcl' (voiced closure) and 'sil' (silence) have not been included.

On the basis of test data, eight phones performed relatively well in terms of validity and were selected for fusion with the baseline system. These phones were /ɛ/, /u/, /ʌ/, /m/, /ɪ/, /n/, /i/ and /ɾ/ as highlighted in Figure 1. Note that the results presented in Figure 1 are those for the test data and not the development data. The baseline result was then calculated for validity and reliability using a common subset of only those utterances containing all eight phones. Due to the large number of possible fused combinations from these eight individual phones, four sub-groups of phones are presented, for brevity. Results show that an improvement over the baseline in validity or reliability, or frequently both, can be gained by fusing combinations of sub-groups {/ɛ/, /u/, /ʌ/, /m/}, {/ɪ/, /n/}, {/i/} and {/ɾ/} with the baseline system, and Table 2 lists some of the more promising results. The best performing overall system was found by fusing the baseline system with /ɪ/ and /n/, which gave an improvement in C_{llr} of 17.02% and improvement in the 95% CI of 5.97% as depicted in Figure 1. Fusing baseline system with /ɛ/, /ʌ/ and /m/, gave the best improvement in the 95% CI of 9.05%, however with only a small improvement in C_{llr} of 2.27%.

3.2. Effect of phone boundary error

An evaluation of phone recognizer endpoint accuracy was made, using TIMIT's manual annotation of phoneme boundaries. A basic algorithmic search was developed to match each of the true labels to the recognized label, such that the overlap between the recognized label and true label was greater than 50% of the duration of the true label, and the recognized and true labels matched. Given that the phone recognizer endpoints contain errors, it is of interest to evaluate the minimum endpoint error achievable if the recognized tokens are adjusted to have shorter or longer duration. An experiment was performed on the TIMIT test database by adjusting (shrinking/expanding) every recognized label by a fixed percentage of its duration. The start and end boundary time differences (recognized versus true labels) at each adjustment were then captured in terms of Root Mean Square Error

(RMSE). The lowest RMSE value for start and end boundaries (i.e. closest to manual annotations of nasal boundaries) were observed to be at 90% duration and 80% duration respectively.

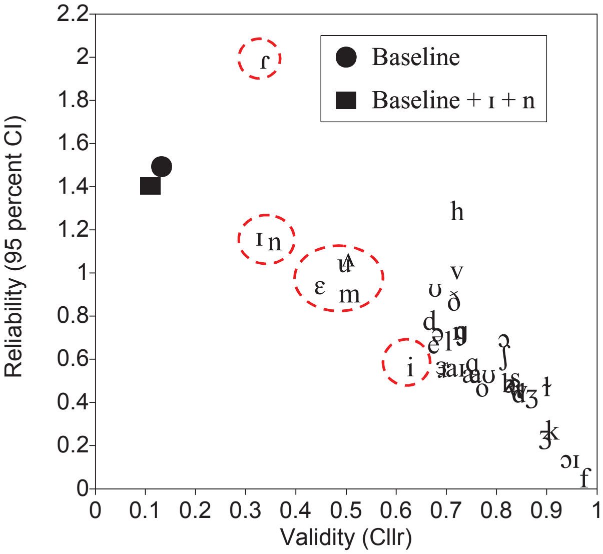

Figure 1: Reliability (95% CI) versus validity (C_{llr}) plot for all individual phones. The baseline system was calculated using a common subset of only those utterances containing all the eight phones highlighted in dashed red.

Table 1: Validity (C_{llr}) and reliability (95% CI) of selected fused systems using the eight best phones with the baseline system.

Fusion	C_{llr}	95% CI
Baseline	0.132	1.492
Baseline + ɛ + ʌ	0.130	1.361
Baseline + ɛ + ʌ + m	0.129	1.357
Baseline + ɪ	0.114	1.420
Baseline + n	0.112	1.438
Baseline + ɪ + n	0.110	1.403
Baseline + i	0.125	1.466

In particular, we were interested to evaluate how the FVC system performs if similarly each of the recognized labels in the NIST database shrinks (or expands) by a fixed amount, denoting different types of phone recognition error. The validity and reliability of the FVC system based on shrinking/expanding of the recognized phone /n/ are shown in Figure 2. Each of the percentages indicates the proportion of token's duration that was used for analysis with 80-90% probably corresponding to the minimum endpoint error. It should be noted that the phone recognizer was trained on TIMIT database (clean read speech), and it was applied on the FVC system that is based on NIST database (telephone speech with noise). Since these are two databases of different nature, we can only speculate that the 80-90% duration may be corresponding to the minimum endpoint error in the FVC system based on NIST database. The results in Figure 2 indicate that shrinking/expanding the token duration creates a trade-off between validity and reliability. In particular, overestimating/underestimating the recognized token's length gives improvement in validity/reliability with a corresponding degradation in reliability/validity. Similar trade-off results were observed for other phones.

In Figure 2, the system with the best validity was observed at the longest token duration of 180%, while the system with the best reliability was observed at the shortest token duration of 50%. It should be noted that 50% corresponds to the lowest possible setting without insufficient frames for adapting the GMMs from tokens of phone /n/, while 180% was arbitrarily determined and could be further extended. An additional step was performed to fuse each of these best results with the baseline system as shown in Table 3. Relative to the fusion of baseline with 100% (i.e. without adjustment), fusion of baseline with the longest duration of 180% gave an improvement in validity while fusion of baseline with the shortest duration of 50% gave an improvement of reliability.

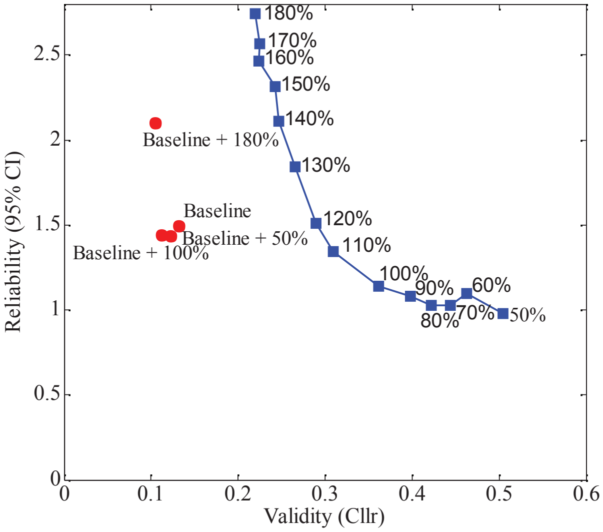

Figure 2: Reliability (95% CI) versus validity (C_{llr}) plot for various amounts of adjustment made to the recognized start and end boundary times of phone /n/ by shrinking/expanding each token as a proportion of its duration (labeled in percentage terms and colored in blue). 100% denotes the original recognized tokens without any adjustment. Fusion combinations based on different adjustment percentages with the baseline were colored in red.

Table 2: Validity (C_{llr}) and reliability (95% CI) of fusion combinations based on different adjustment percentages of FVC system based on phone /n/ with the baseline system.

Fusion	C_{llr}	95% CI
Baseline	0.132	1.492
Baseline + 50%	0.122	1.434
Baseline + 100%	0.112	1.438
Baseline + 180%	0.105	2.101

4. DISCUSSION AND CONCLUSION

Results from our phone-based study have shown that scores from a system that automatically selects and models phones can, when fused with the baseline scores, yield improvement in FVC system validity and reliability, as compared with the baseline system, which employs indiscriminate use of all post-VAD speech frames. In particular, fusion of /ɪ/ and /n/ with a baseline system can yield a substantial increase in system validity and reliability. It may be the case that /n/ has relatively little allophonic variability compared with many other phonemes, which would result in low within-speaker variability and good performance in a FVC system. All systems based on individual phones without fusion with the baseline system had poorer validity than the baseline system (C_{llr} was higher), but (with one exception) their reliability was better (the 95% CI was narrower).

It should be pointed out that the data used for training the UBM in the present study, representing the alternative hypothesis, and for testing the systems, were not completely forensically realistic. Selection of speakers whose recordings were included in these data was restricted to US English male speakers, inevitably containing recordings of many pairs of speakers who sound quite unlike each other and who would never be subjected to forensic voice comparison. In practice recordings should be restricted to those which to a lay person (such as a police officer) sound sufficiently similar to the offender recording that they would think it is appropriate to submit them for forensic analysis. Further, in the current study, all post-VAD frames were used to train the UBM. To be forensically realistic, phonetic tokens within each of the UBM recordings detected on the basis of the phone recognizer should be used in training a phone-dependent UBM.

A trade-off between validity and reliability was observed, as the durations of recognized tokens were varied. A speculative explanation for this may be that by increasing the number of frames within each token, frames from neighbouring speaker-discriminating phones (most likely a vowel if the recognized token is the phone /n/) are included in the analysis, and these phones help to improve system validity. On the contrary, by decreasing the number of frames within each token, we are restricted to strictly only take frames within each token, resulting in high system reliability. An implication is that one may choose to make an adjustment to the recognized segments by effectively increasing/decreasing the number of frames used per token to produce an improvement in validity/reliability, with an accompanying slight degradation in reliability/validity.

Assuming it is possible in a phone recognizer, achieving the minimum endpoint error during token segmentation does not lead to the best validity or reliability for the FVC system based on the NIST database, but rather represents a compromise between these two metrics. As future work, a manually labeled forensic speech database (currently under development) could be used to validate our experimental observation of the trade-off between validity and reliability, and such a database could also be used to build a more forensically realistic phone recognizer.

5. ACKNOWLEDGEMENTS

The writing of this paper was supported financially by NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program

6. REFERENCES

[1] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice*. volume 51, pp. 91-98, 2011.

[2] C. Zhang, G. S. Morrison, and T. Thiruvaran, "Forensic voice comparison using Chinese /iau/," *Proceedings of the17th ICPhS*, Hong Kong, China, pp. 2280–2283, 2011.

[3] I. Magrin-Chagnolleau, J. -F. Bonastre, and F. Bimbot, "Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods," *Proc. EUROSPEECH*, (Madrid, Spain), volume 1, pp. 337–340, 1995.

[4] J. P. Eatock, and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," *Proc. IEEE ICASSP*, pp. 133–136, 1994.

[5] J. -L. Le Floch, C. Montacie, and M. -J. Caraty, "Investigations on speaker characterization from Orphee system techniques," *Proc. IEEE ICASSP*, Adelaide, Australia, volume 1, pp. 149–152, 1994.

[6] T. Bocklet, and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," *Proc. IEEE ICASSP*, Taipei, Taiwan, pp. 4525–4528, 2009.

[7] M. Savic, and S. K. Gupta, "Variable parameter speaker verification system based on hidden Markov modelling," *Proc. IEEE ICASSP*, Albuquerque, New Mexico, USA, 1990.

[8] M. Savic, and J. Sorensen, "Phoneme based speaker verification," *Proc. IEEE ICASSP*, 1992.

[9] N. Ratnayake, M. Savic, and J. Sorensen, "Use of semi-Markov models for speaker-independent phoneme recognition," *Proc. IEEE ICASSP*, San Francisco, California, USA, 1992.

[10] G. S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM)," *Speech Communication*, vol. 53, 2011, 242–256.

[11] K. -F. Lee, and H. -W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 37, no. 11, pp. 1641–1648, 1989.

[12] G.S. Morrison, T. Thiruvaran, and J. Epps, "An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison," *Proceedings of the 13th ASSTA*, Melbourne, Australia, pp. 74–77, 2010.

[13] T. Kinnunen, and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.

[14] N. Brümmer, and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*. vol. 20, no. 2-3, 2006, 230–275.

[15] D. A. van Leeuwen, and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *C. Müller, (ed.), Speaker Classification I: Fundamentals, Features, and Methods*, Heidelberg, Germany: Springer-Verlag, pp. 330-353, 2007.

[16] N. Brümmer, 2005. Tools for fusion and calibration of automatic speaker detection systems. http://niko.brummer.googlepages.com/focal

[17] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.* 10, pp. 19–41, 2000.

[18] T. Becker, M. Jessen, and C. Grigoras, "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models," in *Proc Interspeech*, pp. 1505-1508, 2008.