Speaker verification with elicited speaking styles in the VeriVox project

Karlsson, I.; Banziger Flykt, Tanja; Dankovicová, J.; Johnstone, T.; Lindberg, J.; Melin, H.; Nolan, F.; Scherer, Klaus R.

How to cite

KARLSSON, I. et al. Speaker verification with elicited speaking styles in the VeriVox project. In: Speech Communication, 2000, vol. 31, n° 2/3, p. 121–129. doi: 10.1016/S0167-6393(99)00073-4

This publication URL: https://archive-ouverte.unige.ch/unige:102072

Publication DOI: 10.1016/S0167-6393(99)00073-4

ELSEVIER

Speech Communication 31 (2000) 121–129

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Speaker verification with elicited speaking styles in the VeriVox project

I. Karlsson [a,*], T. Banziger [b], J. Dankovicová [c], T. Johnstone [b], J. Lindberg [a], H. Melin [a], F. Nolan [c], K. Scherer [b]

[a] Department of Speech, Music and Hearing, KTH, S-100 44 Stockholm, Sweden
[b] Department of Psychology, FAPSE, University of Geneva, Geneva, Switzerland
[c] Department of Linguistics, CULD, University of Cambridge, Cambridge, UK

Received 28 August 1998; received in revised form 30 July 1999

Abstract

Some experiments have been carried out to study and compensate for within-speaker variations in speaker verification. To induce speaker variation, a speaking behaviour elicitation software package has been developed. A 50-speaker database with voluntary and involuntary speech variation has been recorded using this software. The database has been used for acoustic analysis as well as for automatic speaker verification (ASV) tests. The voluntary speech variations are used to form an enrolment set for the ASV system. This set is called structured training and is compared to neutral training where only normal speech is used. Both sets contain the same number of utterances. It is found that the ASV system improves its performance when testing on a mixed speaking style test without decreasing the performance of the tests with normal speech. © 2000 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Zum Studium und zur Kompensation von Veränderungen eines Sprechers bei der Sprecherverifikation wurden einige Experimente ausgeführt. Sprecherveränderungen wurden durch eine im Laufe der Arbeit entwickelte Sprechverhaltensstimulanzsoftware erreicht. Eine Datenbank bestehend aus 50 Sprechern und spontane und nicht spontane Sprache wurde mit Hilfe der Software aufgenommen. Diese Datenbank wurde sowohl für die akustische Analyse als auch für automatische Sprecherverifikation (ASV) benutzt. Aus den spontanen Sprachveränderungen wurden ein Registrierungsdatensatz für die ASV erzeugt. Dieser Datensatz wird strukturiertes Training benannt und wird mit neutralem Training, welches nur normale Sprache benutzt, verglichen. Beide Datensätze enthalten die gleiche Anzahl von Sätzen. Die Leistung des ASV Systems verbessert sich für gemischten Sprachstil ohne die Leistung für normale Sprache negativ zu beeinträchtigen. © 2000 Elsevier Science B.V. All rights reserved.

Résumé

Plusieurs expériences ont été produites afin d'étudier et de compenser pour les variations à même un locuteur lors de la vérification du locuteur. Afin de d'encourager les variations à même un locuteur, un logiciel stimulant un tel comportement parlé a été créé. À l'aide de ce logiciel, une base de donnée contenant 50 locuteurs avec variations volontaires

[*] Corresponding author.
E-mail address: inger@speech.kth.se (I. Karlsson).

0167-6393/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(99)00073-4

et involontaires a été enregistrée. Cette base de donnée a été utilisée pour analyse acoustique et pour tests de vérification automatique du locuteur (VAL). Les variations volontaires de la parole sont utilisées comme ensemble d'inscription au système de VAL. Un tel ensemble est appelé entraînement structuré et est comparé à un entraînement neutre consistant seulement de parole normale. Les deux ensembles contiennent le même nombre de phrases. Il est montré qu'en testant sur un style de parole mixte, la performance du VAL système est ameliorée sans que la performance de tests avec parole normale ne dégrade.

Keywords: Speaker verification; Speaker variation; Stressed speech

1. Introduction

The aim of the VeriVox project is to improve the reliability of automatic speaker verification (ASV) by developing novel, phonetically informed methods for coping with the variation in a speaker's voice. It is an attempt to develop a strategy for dealing with factors that significantly alter a person's voice on a short-term basis which according to Furui (1997) is an unsolved problem for automatic speaker verification. In automatic speech recognition (ASR) speech under stress and noise has been analysed and methods for coping with these variations have been developed (Hansen, 1995). For speaker verification some studies of speaker variation over time have been published, see for instance (Furui, 1986).

Up to now, ASV research has treated within-speaker variation as if it were random, but in fact phonetic research reveals that within-speaker variation is highly structured. Different speaking rates, loudness levels, styles, emotional states, and so on, all cause predictable changes in the acoustic speech signal (Murray and Arnott 1993; Junqua 1995). The (long-term) goal is to exploit such known phonetic and phonological regularities to reduce the false rejection rate in ASV without a concomitant rise in the risk of false acceptances. This paper reports on the results achieved during a six-month pilot project, which include a database of various elicited speaking styles, an acoustic analysis of six of the 50 speakers included in the database, and an evaluation with an ASV system.

Our current approach to using phonetic knowledge in an ASV system is called *structured training*. It is the procedure of eliciting different manners of speaking during the enrolment, so that the system becomes familiar with the variation likely to be encountered in that person's voice. One idea behind this approach is to minimise the duration of the enrolment session by including a restricted set of speaking style variations that the enrolee is instructed to produce. Two enrolment sets have been used in SV experiments to test this approach. One set contains neutral speaking style utterances, while the other, referred to as the structured training set, contains utterances with many speaking styles. Both sets have the same number of utterances.

The structured training approach has been tested by comparing it to conventional neutral training using a state of the art HMM-based ASV system developed in the CAVE project (Bimbot et al., 1997). A software package has been implemented for eliciting various types of speaking behaviour and has been used for collecting a database with 50 male Swedish speakers. The database contains speech data to enrol speakers into the system with neutral and with structured training, and further to test the system with a variety of speaking styles. While recordings are made with a high-quality microphone, full bandwidth and a high sampling rate to allow for various acoustic analyses of the speech, the recordings have been transformed to approximate telephone speech quality for the ASV experiments. This is done because many applications of ASV are expected to appear in the context of telephony, and to prevent research from resulting in methods that are applicable to high-quality recordings only.

This paper is organised as follows. The implementation of a method to elicit different speech variations to be used in structured training is described in Section 2, followed by a summary of preliminary acoustic analyses of recordings from six of the 50 speakers in Section 3. The ASV system used in the experiment is described in Section 4, while the database and the experiment itself are described in Sections 5 and 6. Section 7 presents results from the ASV experiment. We conclude by discussing the results and outlining future research plans.

2. Elicitation method

The speech database was recorded using a prototype version of the speaking behaviour eliciting software developed within the project. The software is designed to systematically elicit different types of voluntary and involuntary speech variations. In subsequent trials with an ASV system, the recorded speech samples containing voluntary speech variation are used during the enrolment phase and in test sets, while the recorded samples containing elicited involuntary speech variation are used only in test sets.

Voluntary speech variation is elicited by prompting the user to deliberately speak in a number of different modes, including normal, fast, slow, weak, strong and denasalised speech (pinched nose). For each mode, the user is asked to read aloud six sequences of six digits (2 3 4 5 7 0) in different orders, constructed so that every digit appears following and preceding all other digits. The used digit sequences are listed in Table 1.

The order in which the modes are collected is normal, weak, strong, slow, fast, denasalised and normal again. A limited set of phonetically derived connected phrases is also included in the recordings. These phrases were not used in the experiments reported here. To permit direct comparison among the neutral speech samples and the various induced speech variations, the digits and phrases are standardised across the different conditions.

Table 1
The six digit sequences uttered in the voluntarily elicited speech

0 2 4 7 5 3
5 0 3 4 2 7
2 3 7 0 5 4
3 0 4 5 7 2
4 0 7 3 2 5
7 4 3 5 2 0

The software elicits involuntary variation by means of an interactive module in which users perform a succession of tasks, which cause them to speak normally, faster and louder without being explicitly asked to do so. The tasks include (i) speaking in the presence of two levels of background white noise (administered through headphones), (ii) speaking from memory at an increased rate due to time pressure and (iii) speaking while solving a divided attention logical reasoning and auditory recognition task, with background noise distraction (that is under high cognitive load), eliciting the recording of speech under cognitive stress. Non-directed normal speech samples are also collected as part of this interactive module. All these tasks are designed to elicit the types of involuntary speech variation, which might realistically occur in use of speaker verification systems. This second module (involuntary variation) of the elicitation system uses the same digit sequences as used in the first part (voluntary variation).

After each task, users are asked to evaluate their stress level by clicking on a scale ranging from 0 to 9 with 5 indicating ‘normal’. This self-report of experienced stress level can subsequently be used to control for the effects of unintended induced stress, as well as to evaluate the efficacy of the cognitive stress induction task.

3. Acoustic analysis

Some preliminary acoustic analysis has been performed on the speech database. Segment durations and formant frequencies at vowel mid-points have been measured for six speakers saying 3-0-4-5-7-2 (/tre: nɔl fy:ra fɛm ɧʉ: tvo:/) spoken in seven conditions: Neutral, Loud, Weak, Slow, Denasal (pinched nose) and (Cognitive) Stressed. The six speakers all come from the Stockholm area and their age ranges from 25 to 40 years. One token of each word in each condition is analysed for each speaker, except in the case of Neutral and Cognitive Stressed where two tokens are analysed. For the same six speakers, the durations for other six-digit strings uttered under Low or Loud Background noise and in the ‘Memory under time pressure’ task are compared with the same strings uttered in the Normal condition. Despite the small amount of data a number of trends emerge, some of which are summarised below.

As expected, all segments in Slow are longer, and most in Fast are shorter. Loud and Weak predominantly involve longer segments. Cognitive Stressed and 'Memory under time pressure', see Fig. 1, seem to involve almost consistent shortening of segments while speech against noise conditions shows a strong tendency in the same direction. The self-estimated stress levels indicated by the six speakers are shown in Table 2. As can be seen here, the Cognitive Stressed and the 'Memory under time pressure' are experienced to be about equally stressful.

If each segment's duration is expressed as the percentage change it undergoes as a proportion of the utterance, relative to Neutral, it emerges that a rate change is unevenly distributed over different categories of sound. In Slow there is a clear tendency for the vowels to take up a greater proportion of the lengthening and the consonants less, relative to Neutral; this is also true for Loud. The pattern in Fast is reversed: several consonants take up a greater proportion of the utterance and several vowels take up a smaller proportion.

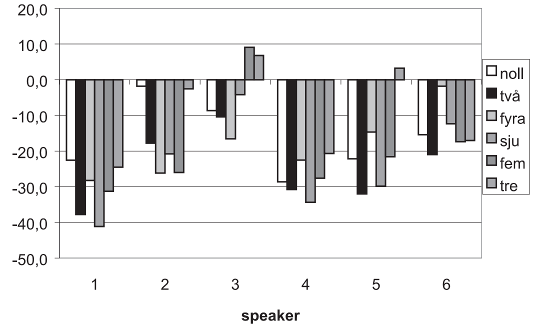

Fig. 1. Duration difference in percent between the same digit sequence in Normal and in 'Memory under time pressure' conditions for six speakers. The digit sequence is [nɔl tvo: fy:ra ɧʉ: fɛm tre:]. The different digits are indicated by different bar patterns. The general trend is towards shorter words in the Memory condition, but the between speaker variation is large.

The first and second formants have also been measured. In Fast, with the exception of /fɛm/, the vowels are mid-centralised. In Fast speech there is perhaps less time for the tongue to achieve peripheral articulations. In Slow the vowels are more peripheral. Cognitive Stressed shows on a smaller scale the (de-peripheralisation) pattern of Fast, while Loud shows in some vowels a pattern of peripheralisation, like Slow. Possibly the changes resulting from Cognitive Stressed (a difficult style to induce) are similar enough to those resulting from Fast that only Fast need be included in the structured training. Full-scale acoustic results will provide a systematic basis for rationalising the 'structured' training.

The styles tested bring about radical restructuring of the temporal and spectral properties of the speech. This gives a clue to why errors arise in the verification process. Given that 'claim' utterances may differ durationally in complex ways, then even if it is possible to 'time-warp' the claim utterances so that they align well with Neutral reference data, the aligned segments (vowels in particular) will match badly in spectral terms.

The database has also been used for more extensive investigations on differences between speaking styles (Karlsson et al., 1998). The results achieved in that study are in accordance with the results discussed above.

4. Speaker verification system

An HMM-based system (Bimbot et al., 1997) is used in the experiments. This system is based on the Hidden Markov model Tool Kit from Entropic (Young et al., 1997). The speech is parameterised using 12 LPC-derived cepstral coefficients

Table 2
Average stress levels for six speakers

After reading in different manners	After talking in noise	After 'Memory under time pressure'	After Cognitive Stressed condition
5	5.5	6.7	6.8

plus energy, with appended delta and acceleration coefficients (totally 39 elements per frame). The LPC-coefficients are derived from an autocorrelation-based LPC analysis of order 16. The analysis window size is 25.6 ms, window shift is 10 ms, pre-emphasis factor is 0.97 and a Hamming window is used. The LPC-derived cepstral parameters are liftered according to formula 1 with liftering parameter L set to 16. Liftering of cepstral parameter number n, with liftering constant L:

$$c'_n = c_n\left(1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right). \qquad (1)$$

The energy is calculated as the log of the signal energy within each analysis window of $N = 204$ speech samples:

$$E = \log\sum_{n=1}^{N} S_n^2. \qquad (2)$$

Each log energy term is floored to 50 dB below the highest energy in the utterance, and normalised to have a fixed maximum energy of 1.0. The energy normalisation is implemented by subtracting the maximum value of the energy in the utterance and adding 1.0 (Young et al., 1997). Delta coefficients are calculated on two preceding and two proceeding frames. Cepstral mean subtraction is applied to static coefficients in order to decrease inter-session variability (Furui, 1994).

Client models have one left-to-right HMM for each digit. The number of states for each digit HMM is set to twice the number of phonemes in the phonetic transcription of the digit (there are between two and four phonemes in Swedish digit words). There are two Gaussian mixture components per state. A world model with the same characteristics as the client models is used for log-likelihood normalisation (Furui, 1994) of the score from a client model. This log-likelihood normalisation is performed on the score obtained for the entire utterance. An inter-word model (silence and garbage) is shared by all client models and the world model. This inter-word model is a five state left-to-right HMM.

When training the world and client models a word boundary segmentation of the training sequences is needed. It is assumed here that an ideal segmentation component is available and this is simulated by using manual segmentations. The model parameters are estimated with the expectation maximization (EM) algorithm, optimizing the maximum likelihood (ML) criterion. During the test session the system automatically makes its own segmentations given the sequence of spoken words, i.e., the system knows which words the client actually said. This is done through a forced alignment using a Viterbi search through the utterance.

The system configuration is one of those that performed well in tests in the CAVE project reported on in (Bimbot et al., 1997). The system implementation used in the experiment is described in the same reference.

5. Database

The database used for true-speaker and false-speaker tests in the ASV experiments has been collected with the speaking behaviour elicitation software described above. It contains a single 30-minute session from each of 50 speakers which includes both enrolment and verification utterances for the speaker. Given that our interest is mainly in speaker variations due to systematic changes in factors like speaking rate and loudness level, we found it reasonable for a first study to use material from a single recording session.

All speakers in the database are male and come from the same (broad) dialect region around Stockholm. The speech material used in the ASV experiments consists of sequences of six digits spoken in Swedish. Each such sequence contains the digits 0, 2, 3, 4, 5 and 7 in various orders.

Recordings have been made in a sound-treated booth with a high-quality head set microphone and a sample rate of 22 kHz. These full bandwidth recordings are used for various acoustic analyses within the project. For a first-order approximation of telephone speech quality in the ASV experiments, recordings have been down-sampled to 8 kHz, band-pass filtered to approximately telephone bandwidth (300–3400 Hz), and finally quantised to 8-bit A-law coding (ITU G.711). Digit sequence boundaries are then marked manually. For the enrolment speech word boundaries are marked.

For training the world and silence models in the ASV system, recordings from 15 male and 15 female speakers in a separate telephone speech database (Melin, 1996) have been used. In this database the selected clients utter 25 five-digit sequences each. All sequences for a client were recorded through one call. In this way the ASV system is set up to work with telephone quality speech and with impostors of both genders.

6. Experiment

The purpose of the experiment is to test the hypothesis that structured training, as defined in Section 1, is helpful in making an ASV system more robust to naturally occurring variations in speaking style, and especially to reduce the false rejection rate at a given false acceptance rate. Two enrolment sets are therefore defined from the first part of the recording session, set A and set B. Set A represents conventional *neutral training* and contains only neutral speech. It serves as a baseline in the experiment. Set B simulates *structured training* and contains equal amounts of all six speaking styles included in the first part of the session (Neutral, Weak, Strong, Slow, Fast and Denasal). Both enrolment sets have equal size and contain 12 six-digit sequences each.

To compare the two enrolment sets, the same batch of tests has been run for each set. For each enrolment set, all 50 speakers are first enrolled as clients in the system, and a set of 31 true-speaker and 49 false-speaker tests per client is then performed. The set of true and false speaker tests is identical for both batches.

Verification utterances for the simulated identity claims are taken from the first and second part of each speaker's session, and each verification test is made with one six-digit utterance. The distribution of the test utterances over speaking styles for the true-speaker claims is shown in Table 3. The full set of true-speaker test utterances is called the *Composite* set and contains a somewhat realistic mix of speaking styles that could appear during use of an ASV system. When analysing results from the experiments, three disjoint subsets of the Composite set will also be referred to, whose composition is also shown in Table 3.

For simulated impostor attempts, one neutral speech utterance from each speaker is selected for an attempt against all other speakers' identities, yielding 49 independent impostor attempts per enrolled client. The same series of impostor attempts are used with each of the partitions of the set of true-speaker tests.

7. Results

The main objective of using structured training is to reduce false rejection rate for a given false

Table 3
The number of true-speaker tests of each elicited speaking style per client[a]

Style	Composite	Neutral	Cognitive Stressed	Other
Neutral	16	16		
Weak	1			1
Strong	1			1
Slow	1			1
Fast	1			1
Denasal	1			1
Noise, weak	1			1
Noise, loud	2			2
Memory, fast	1			1
Cognitive Stressed	6		6	
Total size	31	16	6	9

[a] The *Neutral*, *Cognitive Stressed* and *Other* sets are disjoint subsets of *Composite*.

acceptance rate. Such reductions for various operating points can be read from a detection error trade-off (DET) curve (Martin et al., 1997), and Fig. 2 shows such curves for the two enrolment sets and the *Composite* test set. The reduction in false rejection rate for structured training compared to neutral training, at a fixed false acceptance rate, is the vertical distance between the two corresponding DET curves. If for example, we start with the equal-error-rate point at neutral training, the false rejection rate is reduced from 2.7% to 1.4% when changing to structured training: a 48% reduction in error rate.

Figs. 3 and 4 show DET curves for the individual partitions of the test set. Fig. 3 compares performance with the two enrolment sets for a given partition of the test set. It can be seen that on the *Neutral* test set the system performs equally well with both enrolment sets, while on the *Stressed,* that is Cognitive Stressed, test set performance degrades with structured training. It seems that structured training has not captured the kinds of variations that are associated with elicited cognitive stress. There are some indications in the acoustic analysis of the different speaking styles

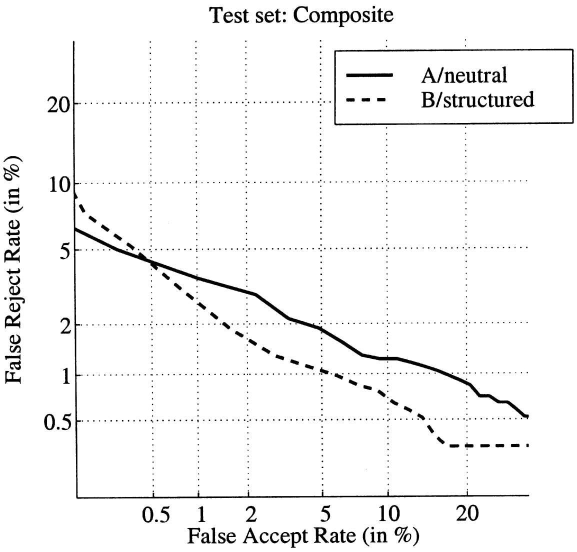

Fig. 2. DET curves for the system with conventional neutral training and with structured training, tested on the Composite test set. The threshold parameter is speaker-independent.

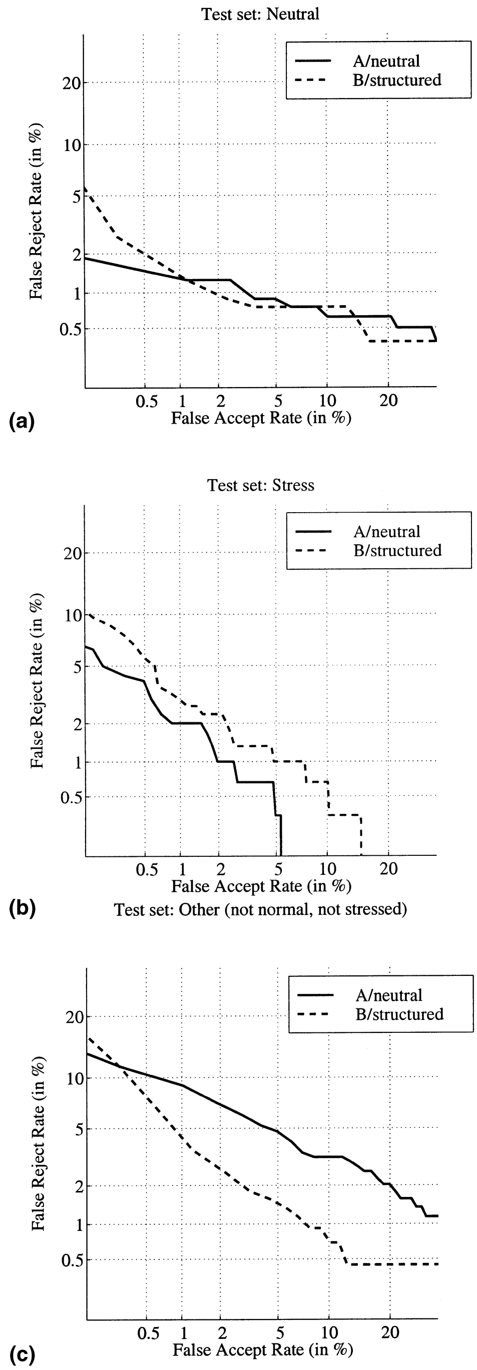

Fig. 3. DET curves for the system with conventional neutral training and with structured training, tested on three subsets of the Composite test set, (a) Neutral test set, (b) Cognitive Stressed test set and (c) Other test set. The threshold parameter is speaker-independent.

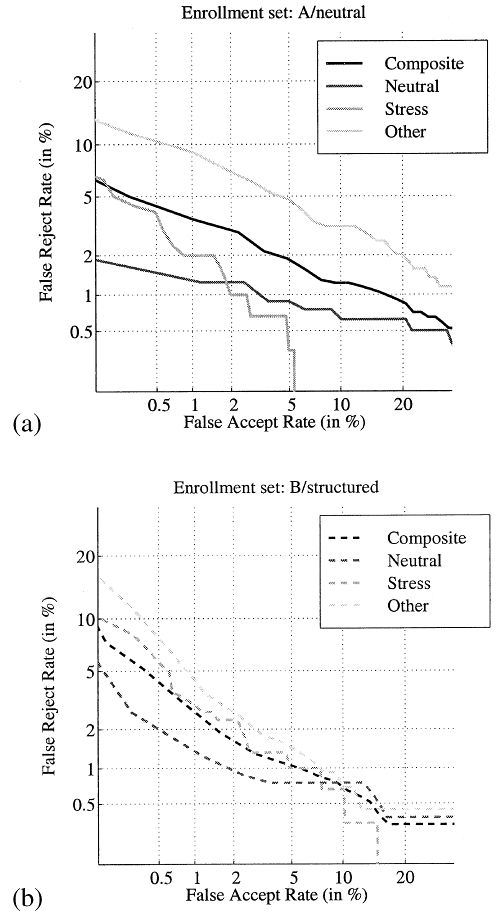

Fig. 4. DET curves for the Composite test set and the three subsets of it, with (a) neutral training, (b) structured training. The threshold parameter is speaker-independent.

mentioned above that Cognitive Stressed speech has much in common with voluntarily elicited Fast speech. Perhaps an enrolment set containing only Neutral and Fast speech would give better results. This remains to be tested. For the *Other* test set there is a significant improvement with structured training. Note that many of the speaking styles included in the *Other*-partition are also included in the structured training.

It would have been interesting to test other groupings of the Composite set as well. It is probable that both the background noise and the time limitation under which some utterances were produced induced stress in the speakers. Due to financial constraints we have not been able to do that so far.

Fig. 4 shows the same DET-curves as those shown in Figs. 2 and 3 but grouped by enrolment set with neutral training in Fig. 4(a) and structured training in Fig. 4(b). It becomes apparent that error rates move apart more with the neutral training than with the structured one.

Table 4 finally, summarises the results in average equal-error-rates with speaker-independent thresholds for each of the two enrolment sets and the different partitions of the test set.

8. Discussion

With neutral training the error rates for the different test sets seem to move apart more than for the same tests when using structured training. This means that the neutrally trained models give a varying performance for the different elicited speaking styles, while the structured training causes a more similar performance for the different speaking styles. This occurs without degrading the performance for the Neutral test set.

With a world model built from neutral training and with structured training of the client model, it is likely that the system would be poor in rejecting impostor attempts with non-neutral speech. The client model is trained to match a broader variation of speech than with neutral training (as long as it is not very well tuned to the particularities of

Table 4
Average (same-sex) equal-error-rate for the two enrolment sets over each of the test sets[a]

Enrolment	Composite	Neutral	Cognitive Stressed	Other
A/Neutral	2.66	1.26	1.67	4.77
B/Structured	1.80	1.26	2.29	2.46

[a] Thresholds are speaker-independent and calculated a posteriori.

the modelled speaker) while the world model is a poor model of impostors speaking in a non-neutral manner. It is therefore important that in a system with structured training, the world (or cohort) model is also created with structured training. In the current experiment, this problem is circumvented by using impostor attempts with neutral speech only.

9. Conclusion

In the VeriVox project, structured training has been tested as a way of making a speaker model in the ASV system familiar with variations in a speaker's voice likely to be encountered in future access attempts. A near halving of the average false rejection rate is demonstrated on a mixed speaking style test set, at no increase in false acceptance rate, and this clearly shows the feasibility of the approach. So far the original ASV system itself has not been modified. A feasible way for further improvement of the system's robustness to variations in speaker style could be to modify the speaker model to better make use of the data seen through structured training. Another possibility is 'guided elicitation', whereby the system guides a client into the right way of speaking in case of a negative outcome from a first verification test.

Acknowledgements

VeriVox was supported by the EU Esprit program. Partners of the consortium are Cambridge University (UK), CNRS (F), KTH (S), Queen Margaret College (UK), University of Dublin (IRL), University Bonn (D), University Geneva (CH), and Ensigma Ltd. (UK). Special thanks to Lennart Nord (KTH) for assistance with the database recording.

References

Bimbot, F., Hutter, H.-P., Jaboulet, C., Koolwaaij, J., Lindberg, J., Pierrot, J.-B., 1997. Speaker verification in the telephone network: Research activities in the CAVE Project. In: Kokkinakis, G., Fakotokis, N., Dermatas E. (Eds.), Proceedings EUROSPEECH '97, Rhodes, 22–25 September 1997, pp. 971–974.

Furui, S., 1986. Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Communication 5, 183–197.

Furui, S., 1994. An overview of speaker recognition technology. In: Proceedings ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 5–7 April 1994, pp. 1–10.

Furui, S., 1997. Recent advances in speaker recognition. In: Bigün, J., Chollet, G., Borgefors, G., (Eds.), Proceedings First International Conference on Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, 12–14 March 1997. Springer, Berlin, pp. 237–252.

Hansen, J. 1995. Analysis and compensation of speech under stress & noise for environmental robustness in speech recognition. In: Proceedings ESCA-NATO Tutorial workshop on Speech under Stress, Lisbon, Portugal, 14–15 September 1995, pp. 91–98.

Junqua, J.-C. 1995. The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. In: Proceedings ESCA-NATO Tutorial Workshop on Speech under Stress, Lisbon, Portugal, 14–15 September 1995, pp. 83–90.

Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K., 1998. Within-speaker variability due to speaking manners. In: Proceedings ICSLP'98, Sydney, 30 November–4 December 1998, pp. 2379–2382.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Kokkinakis, G., Fakotokis, N., Dermatas, E. (Eds.), Proceedings EUROSPEECH '97, Rhodes, 22–25 September 1997, pp. 1895–1898.

Melin, H., 1996. Gandalf – A Swedish telephone speaker verification database. In: Bunnell, H.T., Idsardi, W. (Eds.), Proceedings International Conference on Spoken Language Processing, Philadelphia, 3–6 October 1996, pp. 1954–1957.

Murray, I., Arnott, J., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Amer. 93, 1097–1108.

Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1997. The HTK book, Entropic Research Laboratory Inc. Cambridge, UK.