A Comparative Study of Speaker Verification Systems using the Polycost Database

Nordström T.[*], *Melin H.*[+], *Lindberg J.*[+]

*Telia Research AB, Sweden; Tomas.B.Nordstrom@telia.se
+ KTH, Dep. of Speech, Music and Hearing, Sweden; {melin, lindberg}@speech.kth.se

ABSTRACT

This paper reports on a comparative study of several automatic speaker verification systems using the Polycost database. Polycost is a multi-lingual database with non-native English and mother-tongue speech by subjects from 14 countries. We present results for the first three baseline experiments defined for the database as well as explore the multi-lingual aspects of Polycost in a number of experiments where we compare cross-language and same-language impostor attempts. Our results then lead us to suggest a revised set of baseline experiments.

1. INTRODUCTION

This paper presents some of our findings from a comparative study of several automatic speaker verification (SV) systems. We made the study with the publicly available Polycost speaker verification database [1], which is a multi-lingual database with non-native English and mother-tongue speech by subjects from 14 countries in Europe. For this database a series of four baseline experiments (BE) have been specified [1,2].

The three objectives of the study was

- to compare performance of a number of tools and algorithms on various verification tasks,
- to investigate on the influence from the fact that speakers in a population a) are speaking different languages, or b) have different native languages when speaking English,
- to evaluate the baseline experiments themselves and possibly suggest modified or additional experiments.

Polycost contains around 10 sessions from each of 134 subjects and both English and the subject's mother tongue are spoken in each session. The three baseline experiments with a speaker verification task are: BE1: text-dependent SV on a sentence spoken in English, BE2: SV on (prompted) connected digits spoken in English, BE3: text-independent SV on free speech in the speaker's mother tongue.

For these tasks we have compared up to five different systems. One of the systems is a commercial verifier based on GMMs and composite impostor models of male and female voices for a range of phones. The other systems are two HMM-based ones, one GMM-based and one using second-order statistical measures (SOSM).

2. SYSTEMS

2.1. Feature extraction

Several of the systems described below use the same speech features, or variants thereof. We therefore start by describing the feature extraction part for the HMM-based versions of the GIVES and CAVE systems. The GMM and SOSM versions of GIVES use variants of these features as describe separately for the respective system.

The input signal is pre-emphasized and divided into one 25.6 ms frame each 10 ms and a Hamming window is applied. For each frame a 12-element cepstral vector and an energy term is computed, and those are appended with first and second order deltas. Cepstral mean subtraction is applied to the 13 static coefficients. One of two variants of cepstral vectors are used, MFCC or LPCC. For the MFCC version a 24-channel, FFT-based, mel-warped, log-amplitude filterbank between 300-3400 Hz is followed by a cosine transform. The energy term is the 0'th cepstral coefficient. With LPCC, parameters from a 16-pole linear prediction filter are computed with the autocorrelation method and are transformed to 12-element cepstrum. The energy term is the raw log-energy within each frame of samples, normalized within each utterance to have constant maximum amplitude for every utterance. All cepstral vectors are liftered to equalize their component variances. Total vector size is 39.

2.2. GIVES

GIVES (General Identity Verification System) [8] is a generic platform for speaker verification systems. In this paper we use three different system setups: a text-prompted HMM system; and a GMM and a SOSM-based text-independent system. The first has been tested on BE2, the second on BE3 and the last on all BEs.

LRHMM: A speaker model in this text-prompted system has 10 word-level left-to-right HMMs, one for each digit. Each HMM have two states per phoneme and a mixture of eight Gaussians per state. A non-client model is used for log-likelihood normalization on a per-word basis. Each word score is further divided by the number of frames in the word segment, and finally averaged over words in the utterance. The non-client model is selected individually for each client and each word during enrollment as one of two competing gender-dependent multi-speaker models, with no *a priori* information on the gender of the client. Multi-speaker HMMs are also left-to-right and have the same dimensions as the client HMMs. When training the client model, the best matching multi-speaker model is copied as a seed for the client model. The client model means and mixture weights are then re-estimated (Maximum Likelihood training) from enrollment data while variances and transition probabilities are left untouched. The system is tested with both MFCC and the LPCC-based features.

The system depends on explicit segmentation of the input speech into words during both enrollment and test, the segmentation being produced by a speech recognizer from Nuance [7].

GMM: In this text-independent system client and multi-speaker models are 256-term GMMs. The likelihood ratio is computed in the same way as in the HMM system with one of two multi-speaker models serving as non-client model. The training of the client GMM is also the same. This system has been tested with the MFCC-based features without the deltas. Hence, vector dimension is 13. An energy and zero-crossing rate based end-point detector was used to detect the start and end of an utterance.

SOSM: Client and non-client models are both 12-dimensional covariance matrices computed from MFCC-type cepstral vectors. The MFCC-features are the same as those used with the GMM-based system with the exception that the intermediate filter-bank covers frequency range 0-4000 Hz, which turned out to work better with SOSM in tests on another database. The end-pointer is also the same as with the GMM system. The score for a covariance matrix towards an utterance is computed as one minus the distance, where the distance is a symmetrical-sphericity measure [5]. Client score normalization is done in the same manner as with the GMM-based system with one of two gender-dependent multi-speaker models. The SOSM has been tested on all three baseline experiments. The only change in the system between those tasks is to re-train the non-client models on the corresponding off-line material.

2.3. CAVE

The CAVE generic speaker verification system [3] has been tested on the two text-dependent tasks, BE1 and BE2. The system is based on HTK [4]. In the setup for this paper client models have one left-to-right HMM for each word in BE2 and one single left-to-right HMM for the entire utterance in BE1. Each HMM has two states per phoneme and a mixture of two Gaussians per state. A universal non-client model with the same characteristics as the client model is used for log-likelihood normalization of the score from a client model. This log-likelihood normalization 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 non-client model.

Each HMM is trained separately with Maximum Likelihood training modified to floor variances to the global variance of the Polycost off-line speech material. Client and non-client model HMMs are trained from scratch as opposed to being re-estimated from the non-client model. When training the models a word boundary segmentation of the training sequences is needed. For the digit task (BE2) this segmentation was derived from a speech recognizer from Nuance. For the sentence task (BE1) an energy and zero-crossing rate based end-point detector was used to find the start and end of an utterance. 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 were supposed to say.

This system has been tested with the same MFCC and LPCC-based features as the GIVES (LRHMM) system.

2.4. Nuance Verifier

A commercial verifier from Nuance [7] (version 6.0.4) has been tested on all BEs. It is based on GMMs and composite non-client models of male and female voices for a range of telephone handsets. For all experiments a fixed set of system tuning parameters has been used (the default settings recommended by Nuance). Thus, the the verifier is used more or less ”off-the-shelf”. The only difference in system setup between the different baseline experiments is the choice of non-client model. For BE2 the non-client model was trained on digit material, while for BE1 and BE3 it was trained on a material with general text. Both of these non-client models were delivered with the system. We also tried to re-train (adapt) the non-client models on the off-line material provided with Polycost. Speech features are mel-cepstra similar to the ones used by the GIVES and CAVE systems.

3. DATABASE AND EXPERIMENTS

A set of four baseline experiments (BE) has previously been defined for Polycost to provide a common ground for speaker recognition experiments and to enable cross-site comparisons [1,2]. The three first define speaker verification tasks: BE1 is text-dependent SV with a fixed sentence, the BE2 is digit-prompted with a 10-digit sequence, and BE3 is text-independent. In the third, all subjects speak their mother tongue while in the two first they speak English. The experimental conditions of the baseline experiments were chosen to keep experiments realistic, well-defined and easy to implement. 61 male and 49 female speakers are used both as client and simulated impostors. There are 664 true-speaker tests and 6012 same-sex and 5978 cross-sex impostor attempts in the verification tasks. Enrollment is done with two sessions, except with BE1 where four sessions may be used. 22 speakers have been reserved for training of non-client models. They are one male and one female speaker from 11 different countries.

The current specification of baseline experiment stipulates error-rate figures be computed with a software developed in the CAVE-project. This software computes an individual, *a posteriori* EER threshold for each client, and individual EER are combined to produce several alternative average EERs [2]. Two such figures will be included in our tables below. The first is a same-sex (SS) EER and the second a gender-balanced sex-independent (GBSI) EER which takes into account both same-sex and cross-sex impostor attempts. Since these figures are based on speaker-dependent *a posteriori* thresholds they give very optimistic results as will be seen below, especially when the number of true-speaker tests per client is low. As an alternative we include also a same-sex EER based on a global, speaker-independent (*a posteriori*) threshold.

4. RESULTS

4.1. Performance on baseline experiments

BE1. This baseline experiment uses the English sentence ”Joe took father’s green shoe bench out” as a fixed password sentence shared by all clients, where the same sentence is also available for training of non-client models. This setup simulates a recognition task where all clients share the same password phrase and results will not be directly transferable to a system where each client has their own password phrase.

Table 1 shows results for three systems, where two of them are

inherently text-independent. Only the Cave system is setup to be text-dependent. The Nuance verifier was tested with two versions of its non-client models: first with the original models supplied by Nuance and trained on universal text, and second with the same models adapted to the target sentence with the 22 off-line speakers in Polycost. Table 1 shows a large improvement from adapting the non-client models. This improvement may be partly due to re-training on the target sentence and partly due to inclusion of accents representative of the client population.

		EER (%)		
	Threshold:	*global*	*individual*	
BE	**System**[1]	**SS**	**SS**	**GBSI**
1	Nuance/gmm, retrained ncm	0.62	0.05	0.02
	Nuance/gmm, original ncm	1.53	0.13	0.07
	CAVE/ lrhmm (2,mfcc,w)	3.2	1.0	0.70
	GIVES/sosm (-,mfcc,cg)	6.0	3.2	3.7
2	GIVES/lrhmm (8,lpcc,cg)	0.43	0.08	0.06
	CAVE/lrhmm (2,lpcc,w)	0.52	0.05	0.02
	GIVES/lrhmm (8,mfcc,cg)	1.5	0.30	0.24
	Nuance/gmm, retrained ncm	2.2	0.14	0.08
	Nuance/gmm, original ncm	2.4	0.25	0.12
	CAVE/lrhmm (2,mfcc,w)	2.8	0.80	0.44
	GIVES/sosm (-,mfcc,cg)	6.4	4.0	4.1
3	Nuance/gmm, retrained ncm	11.0	6.3	4.2
	Nuance/gmm, original ncm	11.5	7.2	4.5
	Gives/sosm (-,mfcc,w)	15.1	9.7	10.2
	Gives/gmm (256,mfcc,dcg)	17.1	10.4	8.4

Table 1: System performance on three baseline experiments. In all cases but "Nuance, original ncm" are the non-client models (ncm) trained on material spoken by the 22 Polycost off-line speaker.

BE2. Table 1 also shows results for several systems on BE2. In this experiment, two sessions times four ten-digit utterances are used to enroll clients. A verification test is made with one ten-digit sequence, which is the same for each call and for all clients and is not represented in the enrollment material.

From the table we see that the HMM-based systems perform very well and that the LPCC features outperforms MFCC. For the Nuance verifier we see that the re-training of non-client models on the 22 off-line speakers does not result in a large improvement as was the case in BE1. The only potential benefit from re-training model would be to include representative accents, since the original non-client models were already trained on the target text (digits).

BE3. The lower part of Table 1 shows results on BE3. Enrollment is done with two sessions with an average of 11 seconds of free speech each. A verification test is done on one recording where the subject is asked to say his name, family name, gender, city, country and mother tongue. The average length of these utterances is 5.4 seconds. In this experiment subjects speak their mother tongue rather than English (15% of subjects have English as their mother tongue). We see from the table that in general error rates are many times higher than in BE1 and BE2. The main reason is that BE3 is a text-independent task. As for BE2 the re-training of the Nuance verifier's non-client models does not result in a large improvement.

[1] Parentheses summarize three main features of the HMM-based systems: 1) number of Gaussians per state, 2) speech features and 3) non-client model setup, where 'w' indicates one universal multi-speaker model and 'cg' one of two competing gender-dependent models.

4.2. Language and Accent Dependencies

The Polycost database provides a unique possibility to study language and dialect dependencies in speaker recognition. In BE1 and BE2 subjects often speak a foreign accented English, while in BE3 they speak their own language. Intuitively, it should be easier for a speaker recognition system to tell two speakers of different languages apart than two speakers of the same language. The recognition tasks presented by the baseline experiments should be easier than had the database been mono-lingual with homolingual speakers. We can analyze *how much* easier by computing error-rates for subsets of the BEs with same-language and cross-language impostor attempts only. Table 2 shows results for two subsets of BE2 with three of the better systems from Table 1. Table 2 only consider same-sex tests and then there are 1488 same-language and 5852 cross-language tests. We see that the performance with same-language impostor attempts is considerably worse than with cross-language impostors. This trend is more pronounced with GIVES/lrhmm and LPCC-based features than with the MFCC-based version of it and with Nuance that also uses mel-cepstra based features. The latter system gives 0.39% and 1.9% on the cross-language and same-language subsets of BE1, and 10.3% and 14.2% on BE3.

BE2	**SS-EER (%)**		
Subset	*Gives/lrhmm (8,lpcc,cg)*	*Gives/lrhmm (8,mfcc,cg)*	*Nuance/ gmm*
Baseline experiment	0.43	1.5	2.2
Cross language	0.24	1.2	2.1
Same language	1.4	2.6	2.9

Table 2: Same-sex EERs (with a global threshold) for three systems on BE2 and a number of subsets thereof. 'Language' here refers to the mother tongue of the subject.

4.3. Alternative baseline experiments

From section 4.1 we see that error-rates on BE1 and BE2 can be very low. The number of errors is low and it is difficult to make comparisons between two systems with some statistical significance. Very few speakers contribute to the average error-rates while most speakers show no errors at all.

One possible variation of the current baseline experiments that would increase the number of errors is to reduce the size of enrollment data. Table 3 shows results for two of the systems as in Table 2 with a range of enrollment sets, where we denote *X*s*Y*u an enrollment set with *Y* utterances drawn from *X* sessions. Given the contents of a Polycost session it is possible to make more variations of enrollment sets in BE2 than in BE1 and BE3. The table shows that enrollment set 2s1u, for

example, with a total of two utterances drawn from the two first sessions result in an error-rate which is higher than in the current BEs and more suitable for comparisons. These enrollment sets also correspond better to what is required in a commercial SV application. Figure 1 shows DET-curves for alternative variants of the baseline where all enrollment sets are designed as 2s1u.

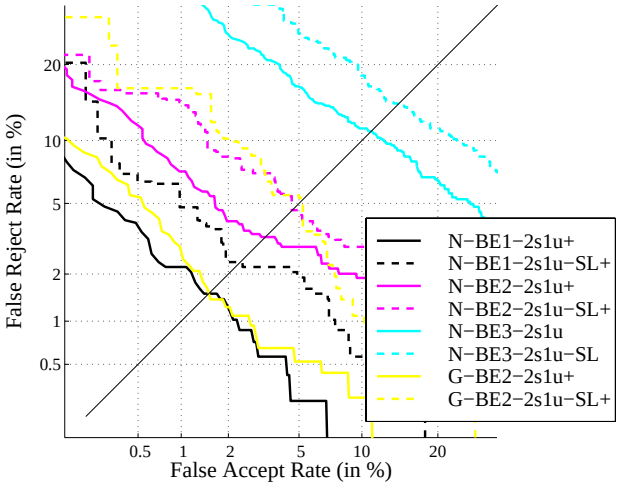

Figure 1: DET-curves [6] based on same-sex impostor attempts for the Nuance verifier[2] (N) and the GIVES/lrhmm system (G) with LPCC-based features. Experiments are the three first BEs modified to have 2s1u enrollment (a '+' in the legend indicates a modified BE enrollment set). For all cases a DET-curve for the subset with same-language impostor tests have been included (SL).

BE	enrollment set	SS-EER % *GIVES/lrhmm (8,lpcc,cg)*	SS-EER % *Nuance/gmm*
1	4s1u (original)	-	0.62
	2s1u	-	1.5 (2.4)
	1s1u	-	3.0
2	2s4u (original)	0.43	2.1
	2s2u	0.55	2.5
	1s4u	1.9	3.9
	2s1u	1.6 (5.1)	3.3 (4.6)
	1s2u	2.7	4.2
	1s1u	3.9	6.6

Table 3: Same-sex EERs (with a global threshold) for the GIVES/lrhmm and Nuance systems on BE2 for different training set sizes. The notation 2s4u means two sessions with four utterances each. The figures within parentheses are the same-language impostor tests EER, cf. Figure 1.

5. DISCUSSION AND CONCLUSIONS

We have presented results for several system on the three baseline experiment on Polycost. The Nuance and GIVES/sosm systems could conveniently be applied directly to all BEs since they are text-independent in their operation. The former system performed well over-all, but was outperformed on BE2 by systems more specialized for a text-dependent task. We also see that cross-language impostor attempts are easier to reject than same-language attempts.

In the presented results, MFCC-based features have performed much worse than LPCC, especially with cross-language impostor attempts. This trend does not hold for some of our experiments on other, monolingual databases. One hypothesis is that since cross-language impostor attempts in Polycost are at the same time "cross-country" attempts (calls originate from different countries), the LPCCs are better suited to recognize where the call come from. We note here that one (main) difference between our LPCC and MFCC is that the latter ignores information in the signal outside 300-3400 Hz while LPCC uses it. Information outside this band may be a cue to differentiate between telephone calls from different countries, where telephone systems are likely to differ more than within countries. If so, this would be a reason to exclude all cross-language impostor attempts from baseline experiments, or comparison may tend to favor systems that are good at call origination recognition in addition to speaker recognition.

Regarding the specification of baseline experiments we therefore suggest to change the specification of baseline experiments in two regards: define all enrollment sets to use one utterance from each of two sessions, and (tentatively) exclude all cross-language impostor attempts. Both changes make the baseline experiments more consistent with each other and more difficult.

[2] Non-client models have been re-trained on the Polycost off-line speakers.

6. REFERENCES

1 Petrovska D., J. Hennebert, H. Melin, D. Genoud, *Polycost: A Telephone-Speech Database for Speaker Recognition*, Proc. RLA2C ("Speaker Recognition and its Commercial and Forensic Applications"), Avignon, France, April 20-23, pp. 211-214. 1998 (or: http://circhp.epfl.ch/polycost)

2 Melin H., J. Lindberg, *Guidelines for experiment on the POLYCOST database*, Proc. Cost-250 workshop, "Application of speaker recognition techniques in telephony", Vigo, Spain, November 1996, pp. 59-70

3 Jaboulet C., J. Koolwaaij, J. Lindberg, J.B. Pierrot, *The CAVE-WP4 Generic Speaker Verification System*, Proc. RLA2C, Avignon, France, April 20-23, pp. 202-205, 1998

4 Young S., J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK BOOK, HTK 2.1 Manual*, 1997

5 Bimbot, Magrin-Chagnolleau, Mathan, *Second order statistical measures for text-independent speaker identification*, Speech Communication 17, 1995, pp. 177-192

6 Martin A., G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, *The DET Curve in Assessment of Detection Task Performance*, Proc. Eurospeech '97, Rhodes, Greece, Sep. 22-25, pp1895-1898, 1997

7 Nuance Communications, 1380 Willow Road, Menlo Park, CA, USA, http://www.nuance.com.

8 Melin H., *On word boundary detection in digit-based speaker verification*, Proc. RLA2C, Avignon, France, April 20-23, pp. 46-49, 1998