THE CRIM SYSTEMS FOR THE 2006 NIST SPEAKER RECOGNITION EVALUATION

Patrick Kenny and Shou-Chun Yin

Centre de recherche informatique de Montréal (CRIM)
Patrick.Kenny@crim.ca, shouchun.yin@crim.ca

1. Introduction

We have submitted 3 gender-dependent factor analysis systems CRIM_1, CRIM_2 and CRIM_3 which differ with respect to the number of speaker factors and in the way that verification scores are normalized and calibrated.

In all cases, UBM's were estimated using the Baum-Welch algorithm, factor loading matrices were estimated with the maximum likelihood and minimum divergence estimation procedures in [1, 2, 3], verification scores were calibrated using logistic regression [4, 5] and the scores produced are intended to be interpreted as log likelihood ratios.

We used the following data sets for development:

- For UBM training: the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and the NIST 2004 evaluation data.

- To train factor loading matrices: the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the NIST 2004 and 2005 evaluation data; and the 2006 auxiliary microphone development data.

- For score calibration: the NIST 2005 evaluation data (in the case of CRIM_1 and CRIM_3) and the Fisher English Corpus, Part 1 (in the case of CRIM_2 only).

The major differences between the three systems are as follows. CRIM_1 and CRIM_2 use identical factor analysis models with 300 speaker factors and 75 channel factors and they differ only with respect to calibration. CRIM_2 was submitted for the core condition only; CRIM_1 was submitted for the 8conv4w-1conv4w, 1conv4w-10sec4w, 10sec4w-10sec4w, 1conv4w-1convmic conditions as well. We did not attempt unsupervised speaker adaptation with either system. The CRIM_3 system has no speaker factors and 75 channel factors. We used it for unsupervised adaptation in the core condition and we also submitted it (without adaptation) for the 8conv4w-1conv4w condition.

We designate CRIM_2 as the primary system for the core condition and and CRIM_1 as the primary system for all of the other conditions which we attempted.

2. The Front End

Where available, ASR transcripts containing time stamps (such as the ctm files provided in the evaluation) were used to suppress silence intervals. In other cases the ISIP voice activity detector is used [6].

Using a 25 ms Hamming window, 12 mel frequency cepstral coefficients together with a log energy feature are calculated every 10 ms. The CPU requirements of this step are $0.002 \times$ RT on a 3.2 GHz Pentium 4 processor where RT stands for Real Time. This 13-dimensional feature vector is subjected to feature warping [7] using a 3 s sliding window ($0.002 \times$ RT). Delta coefficients are then calculated using a 5 frame window giving a 26-dimensional feature vector. This is the feature vector used used for speaker verification in all conditions except for the auxiliary microphone condition.

For the auxiliary microphone data, warped cepstral features were computed just as for telephone speech but they were then mapped onto the telephone domain using a SPLICE transformation [8]. One transformation was trained for each microphone type and gender with the stereo data provided for development using a GMM containing 512 Gaussians. Microphone identification was carried out with another collection of gender-dependent GMM's containing 512 Gaussians; the feature set used for this purpose was 13 dimensional and consisted of the log energy feature normalized over the utterance and raw, unwarped unwarped cepstral coefficients without derivatives. Microphone identification takes about $0.1 \times$ RT. Once the microphone in a recording has been identified, applying the SPLICE transformation takes about $0.1 \times$ RT.

First and second order Baum-Welch statistics are extracted from the non-silence portions of the speech signal using a standard gender dependent universal background model with 2K Gaussians. We regard this as a pre-processing step since we use no information about the speech signal other than that which is encoded in these statistics. This computation takes $0.04 \times$ RT.

3. Factor Analysis

Let C denote the number of mixture components in the UBM and F the dimensionality of the acoustic feature vectors (so that $C = 2048, F = 26$). We assume that if $\boldsymbol{s}$ is the $CF \times 1$ speaker-dependent supervector for a randomly chosen speaker then

$$\boldsymbol{s} \quad = \quad \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{1}$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector, $\boldsymbol{v}$ is a matrix of dimension $CF \times R_S$ where $R_S \ll CF$, $\boldsymbol{d}$ is a $CF \times CF$ diagonal matrix and $\boldsymbol{y}$ and $\boldsymbol{z}$ are random vectors having standard normal distributions. The components of $\boldsymbol{y}$ are *speaker factors*.

We assume further that if $\boldsymbol{M}$ is the $CF \times 1$ speaker- and channel-dependent supervector for a randomly chosen recording of the givenspeaker then

$$\boldsymbol{M} \quad = \quad \boldsymbol{s} + \boldsymbol{u}\boldsymbol{x} \tag{2}$$

where $\boldsymbol{u}$ is a matrix of dimension $CF \times R_C$ where $R_C \ll CF$, and $\boldsymbol{x}$ is a random vector having a standard normal distribution. The components of $\boldsymbol{x}$ are *channel factors*.

The mathematical difficulties with this model arise from the fact that there are two hidden levels: the hidden variables $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ account for supervectors which are themselves hidden — only acoustic feature vectors are observable. However Gaussian assumptions enable all calculations be carried out in closed form.

For all tasks that we attempted other than 1conv4w-1convmic, the hyper-parameters $(\boldsymbol{m}, \boldsymbol{v}, \boldsymbol{d}, \boldsymbol{u}, \boldsymbol{\Sigma})$ were estimated using the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 and 2005 evaluation data. In the 1conv4w-1convmic case, the hyper-parameters $(\boldsymbol{m}, \boldsymbol{v}, \boldsymbol{d})$ were estimated on this data but $(\boldsymbol{u}, \boldsymbol{\Sigma})$ were estimated on the 2006 auxiliary microphone development data.

3.1. Enrolling a target speaker

Given sufficient statistics extracted from an enrollment utterance, we use the prior distribution specified by (1) and (2) to calculate the joint posterior distribution of the hidden variables $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$. Ignoring the channel factors $\boldsymbol{x}$, this gives a posterior distribution on the speaker supervector $\boldsymbol{s}$. For computational reasons we ignore the off-diagonal entries in the posterior covariance of $\boldsymbol{s}$.

In the case of unsupervised adaptation we apply this procedure recursively, obtaining progressively sharper estimates of the speaker's supervector as successive recordings become available. This was the way the CRIM_3 system was deployed in the core condition of the evaluation and in the 8conv4w-1conv4w condition. More information is given in [9].

The computational requirements of this step depends critically on the number of speaker factors and channel factors and not at all on the length of the utterance (given that the signal processing has been carried out). With the configuration $C = 2048, F = 26$, 300 speaker factors and 75 channel factors ~ 300 MB of RAM and 2 min of CPU time are needed; if the number of speaker factors is 0, only 30 MB and 1 s of CPU time are needed. This is the reason why we have only attempted speaker adaptation in the case where the number of speaker factors is 0.

3.2. Likelihood calculations

Suppose we are given a target speaker and a test utterance and we wish to test the null hypothesis that the speaker in the test utterance is different from the target speaker against the alternative hypothesis that the two speakers are the same. Let s be a point estimate of the target speaker's supervector (obtained during enrollment) and let $\boldsymbol{M}$ be the speaker and channel dependent supervector for the test utterance. Under the alternative hypothesis, there is a random vector $\boldsymbol{x}$ such that

$$\boldsymbol{M} \quad = \quad \boldsymbol{s} + \boldsymbol{u}\boldsymbol{x}$$

by (2). If $\boldsymbol{x}$ is known, then so is $\boldsymbol{M}$ so it is a straightforward matter to calculate the (Gaussian) likelihood of the test utterance conditioned on $\boldsymbol{x}$ using the sufficient statistics extracted from the utterance. Thus in order to calculate the (unconditional) likelihood of the test data under the alternative hypothesis, all that is required is to integrate this conditional likelihood against the standard Gaussian kernel. (Recall that x is assumed to have a standard normal distribution.) A closed form expression for this type of integral is given in Proposition 2 of [10]. As explained in [2, 3], a slight correction to the sufficient statistics is all that is required to handle the uncertainty in the point estimate of s that arises from the fact that the enrollment data for the target speaker is of limited duration. (The effect of this correction is most marked in the 10sec4w training condition.)

4. Score Normalization and Calibration

For CRIM_1 and CRIM_2 we applied zt-norm [2, 3] to the likelihood calculated in each verification trial using 200 z-norm utterances and the same number of t-norm speakers. For the t-norm speakers we chose 100 recordings from the 2005 evaluation data and 20 from each of 5 Switchboard databases. We chose the z-norm utterances similarly except in the case of the 1convmic test where the z-norm utterances were taken from the auxiliary microphone development data. In the case of CRIM_3 we used z-norm rather than zt-norm (for reasons explained in [9]) with 100 z-norm utterances taken from the 2005 evaluation data.

For each train/test condition we set the hard decision threshold at 2.29 after applying gender-dependent logistic regression calibration to the normalized verification scores [5]. (This is sometimes referred to as linear calibration since the basic assumption is that the graph of log likelihood ratios versus verification scores is a straight line.) For CRIM_1 and CRIM_3 we estimated the calibration weights for each train/test condition by evaluating on the 2005 test data with factor analysis models having the same numbers of speaker and channel factors that had been trained with pre-2005 data only.[1] Since the 2005 data was incorporated into the training set for the factor analysis models used in CRIM_1 and CRIM_3 this procedure is not ideal (augmenting the factor analysis training set could throw off the calibration).

We also explored a different strategy to estimate the calibration weights for the core condition. We estimated the 2 calibration weights for each gender using a test set consisting of whole conversation sides extracted from the Fisher database. (We did not use Fisher data to train our factor analysis models since we have never found this to be helpful.) For each gender, this test set consisted of 1000 distinct target speakers, 1000 target trials and 10 000 non-target trials. In the female case the equal error rate was 2.0% and the minimum DCF 0.004. The corresponding figures in the male case are 1.8% and 0.005. (Thus Fisher data seems to be much more amenable to speaker recognition than Mixer data. A similar trend is apparent in [11].)

These two stragegies produce quite different estimates of the calibration weights so we averaged the results produced by both to obtain a new 'system', CRIM_2, which we submitted only for the core condition.

5. Computational Cost

The run-time computational requirements for verification of all of our systems are quite modest (assuming a reasonable number of channel factors) *provided* we do not attempt to use them for unsupervised speaker adaptation. Almost all of the computational burden in implementing z-norm is incurred off-line and, as explained in [2, 3], almost all of the run-time computation in a verification trial is common to all t-norm speakers. With the configuration $C = 2048$, $F = 26$ and 75 eigenchannels the CPU time required to match a test utterance with a single hypothesized speaker and 200 t-norm speakers is about 7 s and the memory requirement is ~ 35 MB; if, say, 10 speakers are hypothesized the CPU time per speaker decreases by a factor of 10.

We encountered two types of overhead in unsupervised speaker adaptation. Firstly, every time a speaker model is updated (as in Section 3.1), the mean and standard deviation of the corresponding znorm distribution have to be recalculated. Secondly, it is not easy to take advantge of the fact that several speakers (generally 10 or more) are hypothesized for a given test utterance because speaker models are subject to change over time. For these reasons, our initial implementation of speaker adaptation is very severly I/O bound. It takes almost 2 weeks of elapsed time to perform 6 000 trials (we are only getting about 2% of the CPU); the memory requirements are no different from those of unsupervised adaptation.

6. References

[1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms." [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in

[1] But in the case of the 1convmic test we used the auxiliary microphone development data to estimate both the channel factor loading matrix and the calibration weights.

speaker recognition," to appear in *IEEE Trans. Audio Speech and Language Processing*. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[3] ——, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP 2006*, Toulouse, France, May 2006.

[4] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[5] N. Brümmer. (2005) Tools for fusion and calibration of automatic speaker detection systems. [Online]. Available: http://www.dsp.sun.ac.za/ nbrummer/focal/index.htm

[6] Institute for Signal and Information Processing, Mississippi State University. [Online]. Available: http://www.isip.msstate.edu/projects/speech/software/legacy/index.html

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, June 2001.

[8] L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 477–480, June 2005. [Online]. Available: http://research.microsoft.com/srg/papers/2005-deng-spl.pdf

[9] S.-C. Yin, P. Kenny, and R. Rose, "Experiments in speaker adaptation for factor analysis based speaker verification," in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006.

[10] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[11] L. Ferrar, K. Sönmez, and S. Kajarekar, "Class-dependent score combination for speaker recognition," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.