AN EFFICIENT INTEGRATED GENDER DETECTION SCHEME AND TIME MEDIATED AVERAGING OF GENDER DEPENDENT ACOUSTIC MODELS

Peder A. Olsen and Satya Dharanipragada

IBM, T. J. Watson Research Center
134 and Taconic Parkway
Yorktown Heights, NY 10598
{*pederao,satya*}*@us.ibm.com*

ABSTRACT

This paper discusses building gender dependent gaussian mixture models (GMMs) and how to integrate these with an efficient gender detection scheme. Gender specific acoustic models of half the size of a corresponding gender independent acoustic model substantially outperform the larger gender independent acoustic models. With perfect gender detection, gender dependent modeling should therefore yield higher recognition accuracy without consuming more memory. Furthermore, as certain phonemes are inherently gender independent (e.g. silence) much of the male and female specific acoustic models can be shared. This paper proposes how to discover which phonemes are inherently similar for male and female speakers and how to efficiently share this information between gender dependent GMMs. A highly accurate gender detection scheme is suggested that takes advantage of computations inherently done in the speech recognizer to detect the gender at a computational cost that is negligible. By making the gender assignment probabilistic an increase in word error rate (WER) seen for erroneously gender labeled speakers is avoided. The method of gender detection and probabilistic use of gender is novel and should be of interest beyond mere gender detection. The only requirement for the method to work is that the training data be appropriately labeled.

1. INTRODUCTION

Gender specific models are known to yield improved accuracy over gender independent models and have previously been considered extensively in the literature. The most typical use is a two-pass approach where in the first pass a gender-detection scheme is used to detect the gender of a speaker and in the second pass the speech is recognized with the corresponding gender specific acoustic model. See [1] for an example of sophisticated use of gender information. Other references are [2, 3].

The experiments described in this paper was performed on an IBM internal database, citeDeligne:02,olsen-icassp2002. The baseline acoustic model consisted of a standard 39 dimensional FFT-based MFCC frontend (13 dimensional cepstral vectors and corresponding Δ and $\Delta\Delta$ cepstral vectors spliced together). Digits are modeled by defining word specific digit phonemes, yielding word models for digits. In total 680 word internal triphones are used to model acoustic context and the gaussian mixture models used to model the individual allophones consisted of a total of 10253 gaussians. The number of gaussians assigned to each allophone was determined using the Bayesian Information Criterion as described in [6]. The database used for training was well balanced between the genders. It consisted of a total of 462388 utterances out of which 228693 coresponded to female speakers and 233695 corresponded to male speakers. The test set was similarly well balanced with a total of 73743 words out of which 36241 words were uttered by female speakers and 37502 by male speakers.

2. COMPARISON OF GENDER DEPENDENT AND GENDER INDEPENDENT MODELS

By a male, female or gender dependent GMM we mean a GMM built from the portion of the training data uttered by a speaker of that specific gender. Since a gender dependent GMM is built from roughly half of the training data, it is strictly speaking not obvious that a gender dependent model will outperform a gender independent model built from the entire training data. One test of the usefulness of gender is that a gender dependent GMM of the same size as the gender independent GMM should outperform the gender independent model on test data for speakers of that same gender. Table 1 shows performance on diagonal covariance GMMs corresponding to the gender dependent and gender independent models each with a total of 10253 gaussians. Also, listed in Table 1 is the performance for MLLT (semi-tied covariance) gaussians, [7, 8]. Two points are worth noting in the table. Firstly, that the oracle yields a 29.7% and 28.5% relative improvement in the error rate over respecively the baseline diagonal or MLLT model. Secondly, the cross-gender performance, i.e. a female GMM decoding male speech or a male GMM decoding female speech, is dramatically worse than the gender independent performance. The first point implies that there is a lot of room from improvement. The second point implies that a gender classification error will be very costly. On the other hand, the high cross gender classification error indicates that the models are quite different thus one may suspect that gender classification will be a simple task.

In our target application memory is severely constrained.

Test Gender	Gender of training data both	female	male	oracle
	diagonal GMMs			
both	3.34%	6.22%	6.52%	2.41%
female	4.40%	2.90%	11.27%	2.90%
male	2.32%	9.42%	1.93%	1.93%
	MLLT GMMs			
both	2.95%	5.95%	6.52%	2.11%
female	3.69%	2.48%	11.10%	2.48%
male	2.24%	9.30%	1.76%	1.79%

Table 1. Word error rates broken down on gender for 10K gender dependent and gender independent GMMs

Thus, it is out of the question that the number of gaussians can be doubled even if only half of the gaussians is used once the gender has been determined. Table 2 shows the performance for male and female models that consists of less than half as many gaussians, i.e. 5034 gaussians. The relative improvement for the oracle model is now 19.8% and 19.0% respectively for the diagonal and MLLT models.

Test Gender	Gender of training data 10K, both	female	male	oracle
	diagonal GMMs			
both	3.34%	6.75%	7.27%	2.75%
female	4.40%	3.45%	12.66%	3.45%
male	2.32%	9.93%	2.06%	2.06%
	MLLT GMMs			
both	2.95%	6.61%	7.01%	2.39%
female	3.69%	2.89%	12.29%	2.89%
male	2.24%	10.21%	1.90%	1.90%

Table 2. Word error rates broken down on gender for 5K gender dependent and 10K gender independent GMMs

3. USING GENDER INFORMATION PROBABILISTICALLY

The improvement in the oracle model for the merged 5K gender models are noticeably smaller than for the 10K models, but still substantial. When using a gender detection scheme to detect gender there will inevitably be errors, especially at times of gender changes. As the crossgender performance is very poor, a scheme with a less dramatic deterioration in the word error rate would be desirable. The gender independent 10K GMMs is of course such a model. Table 3 shows the performance for three different interpolation values for the diagonal covariance GMMs. Note that the performance of the model where the male and female GMMs are equally interpolated is only slightly worse than the performance of the gender independent models. What this means is that if it is difficult to assess the gender one can simply use the model $0.5 * \mathrm{GMM_f} + 0.5 * \mathrm{GMM_m}$ at little cost in accuracy.

Test Gender	0.5*$\mathrm{GMM_f}$ +0.5*$\mathrm{GMM_m}$	0.8*$\mathrm{GMM_f}$ +0.2*$\mathrm{GMM_m}$	$\mathrm{GMM_f}$
both	3.51%	3.44%	6.75%
female	4.60%	4.04%	3.45%
male	2.46%	2.87%	9.93%

Table 3. Word error rates for interpolated gender dependent diagonal GMMs.

Let p_f and p_m, $p_\mathrm{f} + p_\mathrm{m} = 1$ represent how certain we are that speech originated from a speaker of a particular gender. If the only acoustics observed from a speaker is a single frame $\mathbf{x}_t$ the best estimate for p_f is the aposteriori gender probability

$$\gamma_{\mathrm{f},t} = \frac{\sum_{g \in \mathcal{F}} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{g \in \mathcal{G}} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)},$$

where $\mathcal{G}$ is the collection of all gaussians and $\mathcal{F}$ and $\mathcal{M}$ are the collection of gaussians corresponding to male and female speakers. With more speech the estimate can of course be improved. With frames $\mathbf{x}_1, \ldots, \mathbf{x}_T$ a reasonable estimate for p_f is simply

$$p_\mathrm{f}(T) = \frac{1}{T} \sum_{t=1}^{T} \gamma_{\mathrm{f},t}.$$

The problem with this estimate is that it does not easily allow detection of a change of speaker. One possible method to fix this is to not use all previous frames, but to create a moving window, i.e.

$$p_\mathrm{f}(T) = \frac{1}{n} \sum_{t=T-n}^{T} \gamma_{\mathrm{f},t}.$$

A tiny drawback to this strategy is that it requires the memorization of the previous $n-1$ values of $\gamma_{\mathrm{f}t}$. Also, this strategy weights each previous sample equally. Intuitively the most current acoustic information should carry more weight than the older acoustic information. A probability distribution solving these two problems is the discrete geometric probability distribution $q_i = (1-\alpha)\alpha^i$, $i = 0, 1, \ldots$. With this distribution we define $p_\mathrm{f}(T)$ by

$$p_\mathrm{f}(T) = \sum_{t=0}^{\infty} q_t \gamma_{\mathrm{f},T-t}.$$

This quantity can now be efficiently computed by the formula

$$p_\mathrm{f}(T) = \alpha * p_\mathrm{f}(T-1) + (1-\alpha) * \gamma_{\mathrm{f},T-t},$$

requiring only the memorization of $p_\mathrm{f}(T-1)$[1]. The mean of q_i is $\alpha/(1-\alpha)$ which can be interpreted as the effective window size when using the weights q_i. In the speech recognizer cepstral vectors are computed every 15ms and α was chosen so that $\alpha/(1-\alpha) = 100$. Thus, the effective gender switching time for $p_\mathrm{f}(T) * \mathrm{GMM_f} + p_\mathrm{m}(T)\mathrm{GMM_m}$ is of the order of 1.5 seconds. The decoding result with the acoustic model

[1] The only distributions with the "no memory property" is the geometric distribution and for continuous distributions the exponential distribution

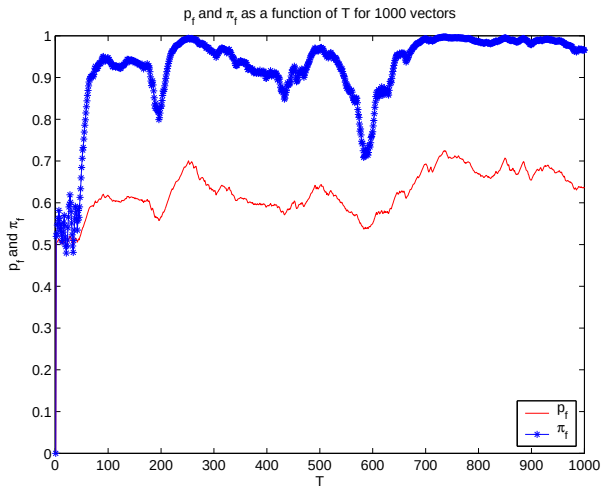

Fig. 1. Graph of $p_{\mathrm{f}}(T)$ and $\pi_{\mathrm{f}}(T)$ for the first 1000 cepstral vectors uttered by a female speaker.

$p_{\mathrm{f}}(T) * \mathrm{GMM_f} + p_{\mathrm{m}}(T)\mathrm{GMM_m}$ is given in Table 4. This acoustic model did not capture much of the gain inherently available in the oracle model. Detailed analysis shows that this is due to $p_{\mathrm{f}}(T)$ and $p_{\mathrm{m}}(T)$ being very close to 0.5. This could mean that $p_{\mathrm{f}}(T)$ is not a good predictor that speech originated from a female speaker, but luckily this is not so. $p_{\mathrm{f}}(T)$ tend indeed to be greater than 0.5 for female speech as can be seen in Fig. 1. The cure that is needed is a "sharpening" of the aposteriory probabilities $p_{\mathrm{f}}(T)$ and $p_{\mathrm{m}}(T)$. Introduce the boosted gender detection probabilities $\pi_{\mathrm{f}}(T)$ and $\pi_{\mathrm{m}}(T)$ by

$$\pi_{\mathrm{f}}(T) = \frac{p_{\mathrm{f}}(T)^{\beta}}{p_{\mathrm{f}}(T)^{\beta} + p_{\mathrm{m}}(T)^{\beta}}. \tag{1}$$

The larger β the sharper the $\pi_{\mathrm{f}}(T), \pi_{\mathrm{m}}(T)$ probabilities become. Table 4 shows results for decoding with the model $\pi_{\mathrm{f}}(T)\mathrm{GMM_f} + \pi_{\mathrm{m}}(T)\mathrm{GMM_m}$ for $\beta = 6$. As can be seen almost all of the gain in the oracle model, which has an error rate of 2.75%, is captured by this acoustic model.

Test Gender	baseline	$p_{\mathrm{f}} * \mathrm{GMM_f}$ $+p_{\mathrm{m}} * \mathrm{GMM_m}$	$\pi_{\mathrm{f}} * \mathrm{GMM_f}$ $+\pi_{\mathrm{m}} * \mathrm{GMM_m}$
both	3.34%	3.29%	2.88%
female	4.40%	4.26%	3.61%
male	2.32%	2.34%	2.18%

Table 4. Word error rates for time mediated averaging of the gender dependent diagonal GMMs.

4. SHARING OF GAUSSIANS BETWEEN GENDER DEPENDENT MODELS

It is clear that silence is inherently gender independent and thus many of the gaussians modeling silence are bound to be unnecessary. Possibly even some of the other phonemes are inherently not different under gender variations too. If we share the gaussians for the sounds that are inherently gender independent we may be able to squeeze out some of the difference between the 10K oracle and 5K oracle models. To measure the difference between two acoustic models for a phoneme X we use the Kullback Leibler divergence

$$D(f||m) = \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{m(x)}\right) f(x)dx. \tag{2}$$

If $f = \mathrm{GMM_f}(X)$ and $m = \mathrm{GMM_m}(X)$ consists of a single gaussian (2) can be computed exactly. Otherwise, the distance must be computed numerically. Monte Carlo estimation can be used to compute the integral in the general case. Let $\{x_i\}_{i=1}^n$ be n samples from the distribution $f(x)$, then

$$\int_{\mathbb{R}^d} f(x) \log m(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \log m(x_i).$$

Using the Kullback Leibler distance we can now decide which phonemes vary little between the genders. To take advantage of this we built gender dependent acoustic models with 6.3K gaussians and gender independent models with 7K gaussians. To combine these we computed the Kullback Leibler distance between all context dependent phonemes and sorted these. We can afford a total of 10K gaussians. Combining the 6.3K male and female acoustic models gives a total of 12.6K gaussians. To reduce the number of gaussians we sort the context dependent phonemes according to the Kullback Leibler distance and replace with gaussians from the gender independent gaussians starting with the smallest distance first. When the number comes below 10K we stop. Table 5 shows the decoding results. Table 6 shows the list of phonemes with smallest and largest Kullback Leibler distance.

Test Gender	baseline	$\pi_{\mathrm{f}} * \mathrm{GMM_f}$ $+\pi_{\mathrm{m}} * \mathrm{GMM_m}$
both	3.34%	2.80%
female	4.40%	3.55%
male	2.32%	2.07%

Table 5. Word error rates for time mediated averaging of the gender dependent diagonal GMMs with shared gaussians.

$D(f\|\|g)$	phoneme	$D(f\|\|g)$	phoneme
0.5059	PD_3	18.3031	OW_1
0.5322	F_1	16.8553	EH_1
0.5626	F_2	16.6865	ER_3
0.6652	F_3	16.3531	EY_3
0.7608	H_1	16.3488	EH_1
0.7662	SIL_1	16.3469	EH_2

Table 6. Top few context dependent phonemes (allophones) with largest and smallest Kullback Leibler distance.

4.1. Fast gaussian evaluation

The previous experiments are a bit unrealistic in that not all the gaussians are evaluated for every frame in a real time speech recognizer. In computing $\pi_{\mathrm{f}}(T)$ we will only have a small set of gaussians that are evaluated for each frame. This may possibly lead to poorer performance in the gender labeling. Table 7 shows the results for the gaussian models considered in Table 5. As most speech recognizers are highly optimized with respect to computational cost even the computation of $\pi_{\mathrm{f}}(T)$ and $\pi_{\mathrm{m}}(T)$ can be prohibitively expensive. One way to save on computation is to further reduce the number of gaussians available in the computation of $\pi_{\mathrm{f}}(T)$ and $\pi_{\mathrm{m}}(T)$. In the extreme where we only keep one gaussian the quantity $\gamma_{\mathrm{f},t}$ is either 0 or 1. We will denote this case by $\hat{\pi}_f(T)$ and $\hat{\pi}_m(T)$. The computation of $\hat{\pi}_f(T)$ now merely corresponds to simple counting and the evaluation of (1) and as can be seen in Table 7 the word error rate actually improves slightly.

Test Gender	baseline	$\pi_{\mathrm{f}} * \mathrm{GMM}_{\mathrm{f}}$ $+\pi_{\mathrm{m}} * \mathrm{GMM}_{\mathrm{m}}$	$\hat{\pi}_f * \mathrm{GMM}_{\mathrm{f}}$ $+\hat{\pi}_m * \mathrm{GMM}_{\mathrm{m}}$
both	3.72%	3.23%	3.21%
female	4.73%	3.87%	3.82%
male	2.73%	2.61%	2.62%

Table 7. Word error rates for decodings with fast hierarchical evalutation of GMMs and a fast gender detection scheme.

5. RETRAINING GENDER AVERAGED ACOUSTIC MODELS

Just as we could merge gaussians to share the common structure in the acoustic models we could imagine letting the EM algorithm automatically discover such structure. If we force $\pi_{\mathrm{f}}(T)$ and $\pi_{\mathrm{m}}(T)$ take on the values 0 or 1 according to the gender of the speaker in the training data the new models will not differ from the current models. Similarly experiments showed that using the values $\pi_{\mathrm{f}}(T)$ described for decoding does not yield any gains either. However if we fix $\pi_{\mathrm{f}}(T)$ and $\pi_{\mathrm{m}}(T)$ to 0.5 and follow the following procedure. First collect statistics on the female training data with the composite model, but update only the gaussians corresponding to the female speakers. Then repeat for male speakers and iterate this procedure. The improvements after several iterations are shown in Table 8.

Test Gender	$\pi_{\mathrm{f}} * \mathrm{GMM}_{\mathrm{f}}$ $+\pi_{\mathrm{m}} * \mathrm{GMM}_{\mathrm{m}}$	retrained
both	3.23%	3.09%
female	3.87%	3.63%
male	2.61%	2.57%

Table 8. Word error rates for retrained averaged gender dependent GMMs.

6. CONCLUSION

This paper describes a technique that takes advantage of gender information in the training data and shows how to squeeze out the most of this information. By using a probability value to average male and female GMMs the dramatic deterioration in cross gender decoding performance is avoided. Finally, the computation needed is negligible and no extra memory is needed to see a substantial drop in the word error rates.

7. REFERENCES

[1] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, and S. J. Young, "Experiments in broadcast news transcription," in *Proceedings of ICASSP*, May 1998.

[2] Francis Kubala, Hubert Jin, Spyros Matsoukas, Long Nguyen, Rich Schwartz, and John Makhoul, "The 1996 BBN Byblos Hub-4 transcription system," in *Proceedings of the Speech Recognition Workshop*, Chantilly, Virginia, February 2-5 1997, DARPA, pp. 90–93.

[3] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and Michèle Jardino, "The LIMSI 1998 Hub-4E transcription system," in *Proceedings of the DARPA Broadcast News Workshop*, Herndon, Virginia, February 28 - March 3 1999, DARPA, pp. 99–104.

[4] Sabine Deligne, Satya Dharanipragada, Ramesh Gopinath, Benoit Maison, Peder Olsen, and Harry Printz, "A robust high accuracy speech recognition system for mobile applications," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 551–561, November 2002.

[5] P. Olsen and R. A. Gopinath, "Modeling inverse covariance matrices by basis expansion," in *ICASSP*, Orlando, Florida, 2002, submitted.

[6] S. S. Chen and R. A. Gopinath, "Model selection in acoustic modeling," in *Eurospeech*, Budapest, Hungary, Spetember 1999.

[7] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions in Speech and Audio Processing*, 1999.

[8] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proceedings of ICASSP*, Seattle, USA, 1998, vol. II, pp. 661–664.