SELF-ADAPTATION USING EIGENVOICES FOR LARGE-VOCABULARY CONTINUOUS SPEECH RECOGNITION

Patrick Nguyen[1,2], *Luca Rigazio*[1], *Roland Kuhn*[1], *Jean-Claude Junqua*[1] *and Christian Wellekens*[2]

[1] Panasonic Speech Technology Laboratory
3888 State Street, Suite 202,
Santa Barbara, CA 93105, U.S.A
email: {nguyen, rigazio, kuhn, jcj}@research.panasonic.com

[2] Institut Eurécom
2229 Route des Crêtes, B.P. 193
06904 Sophia-Antipolis, France
email: welleken@eurecom.fr

ABSTRACT

In this paper, we present the application of eigenvoices to self-adaptation. This adaptation algorithm happens to be rather well-suited for such a task. First, it is an extremely fast adaptation algorithm, and thus well tailored to work for very short amounts of adaptation data. It is also believed to be rather more tolerant of errorful recognition. A third property is the explicit aim to reduce the dimensionality that translates into compact computation of the likelihood. This can be exploited as an embedded confidence measure to minimize the impact of errors in the transcription.

Our experiments were carried out on the Wall Street Journal evaluation task (WSJ). We reduced our word error rate (WER) by one percent absolute to 9.7%.

1. INTRODUCTION

With the advances in recent research, the availability of large speech corpora, and the growing computational capabilities, large-vocabulary speech recognition has become affordable. However, while we are able to build models for, say, voice dictation that comprise about a million of parameters, model adaptation becomes increasingly difficult: the amount of speech available for a specific speaker is limited to at most half an hour, while the complexity (the number of degrees of freedom) of the speech recognizer can grow arbitrarily.

Sometimes, it just so happens that we have absolutely no preliminary speech from the speaker. In that case, we have to perform speaker adaptation for each utterance "on the fly". This is the purpose of self-adaptation: given general-purpose, speaker-independent models, how to incorporate partial knowledge of the speaker from the current speech to more aptly recognize that same speech?

In that scenario, we can readily state the desired properties the adaptation scheme. It needs to be *rapid*, in the sense that we have a only a rather modest fraction of speech available for our purposes. Also, if it is based on partial estimation of what is being said, it must minimize that dependency: to be *robust* to erroneous hypotheses.

2. SELF-ADAPTATION

Self-adaptation is the process by which one adapts models on the same utterance as the one we are currently trying to recognize. Typically, the decoder proceeds in two (or more) passes. The first pass would employ rather coarse models to narrow down the search space to a size that is affordable in subsequent passes. Information is added between passes: vocal tract length, trigrams, cross-word modelling, etc. Statistics or word alignment that were generated in the first pass can be almost readily used for adaptation. A popular, ubiquitous adaptation technique is MLLR (Maxmimum Likelihood Linear Regression [1]). It is commonplace to apply one or more iterations of MLLR adaptation at that stage. Channel mismatch as well as speaker mismatch are thought to be solved in the process. Note that the use of indirect parameters in the adaptation process implies that errorful transcriptions are averaged in with correct ones. The impact of errors on adaptation are present in all adapted parameters. However, performance of self-adaptation is a function of the overall performance of the speaker-independent model.

In this paper, we argue that another adaptation technique, called eigenvoices [2], may be considered as a competitive alternative to MLLR to perform speaker adaptation between passes. First, we summarize the eigenvoices and its associated notations. Then we show how to compute the gain in likelihood of an observation using eigenvoices. In the next step we proceed to explain how to use this to minimize errors in the estimation of the adaptation parameters. Finally, results are presented with a short discussion.

3. EIGENVOICES

Eigenvoices is an adaptation algorithm that employs *a priori* knowledge about the speaker model space [2]. The gist is to create speaker adapted models using the training database, and observe the distributions of the HMM model parameters, to deduce a compact, low-dimensional representation of what a speaker-adapted model is expected to look like.

Model parameters of all Hidden Markov Models (HMMs) of one speaker-adapted model are constrained to lie in a linear vector space, called speaker space. We only consider adaptation of the mean vectors. Let $\bar{\mu}(e)$ be the basis vectors that span the speaker space. They are called *eigenvoices*. There is only a small number E of eigenvoices, typically in the range of 1–100. The rationale behind this term is that they are discovered using an eigen decomposition of the whole set of speakers in the training database. If $\bar{\mu}_m(e)$ is the e-th component that corresponds to the m-th gaussian distribution in the system, then for all m, we can write

$$\mu_m = \sum_{e=1}^{E} w_e \bar{\mu}_m(e)$$

where $\{w_e\}$ represents the location of the speaker in the speaker space. We define $w = [w_1, ..., w_E]^T$.

Given this constraint, given incoming speech O and our *a priori* knowledge $\bar{\mu}(e)$, we find the maximum-likelihood (ML) eigenvoice decomposition (MLED) for that observation, which is done by iteratively optimizing the quadratic exponent function:

$$Q = -\sum_{t,m} \gamma_m(t)(\mu_m - o_t)^T C_m^{-1}(\mu_m - o_t)$$

where o_t is the observation vector at time t, $\gamma_m(t)$ is the posterior probability that the distribution m produced o_t at that time, and C_m^{-1} is the precision matrix of that distribution. The MLED estimation is then equivalent to solving the linear system

$$\sum_{t,m,j} \gamma_m(t) w_j \bar{\mu}_m^T(j) C_m^{-1} \bar{\mu}_m(e) = \sum_{t,m} \gamma_m(t) \bar{\mu}_m^T(e) C_m^{-1} o_t,$$

(with $e = 1...E$). In the next section, we show how to compute the likelihood for concatenations of speech segments.

4. COMPACT SUFFICIENT STATISTICS FOR THE LIKELIHOOD

4.1. Definition

In this section, we find statistics required to compute the likelihood. The idea is that this set of variables $\mathcal{S}$ will enable us to compute the likelihood of a segment of speech with respect to some eigenvalues. That is, a segment of speech can be summarized compactly in $\mathcal{S}$ as far as the computation of likelihood of eigenvoice-adapted models is concerned. Define θ to be the completion data in the Expectation-Maximization (EM) algorithm, i.e. the state segmentation in the Viterbi approximation. It is quite trivial to see that the likelihood of an observation O satisfies

$$-\log p(O, \theta|w) \propto \sum_{t,m} \gamma_m(t) \Bigg[\sum_{e,j} w_e w_j \bar{\mu}_m^T(e) C_m^{-1} \bar{\mu}_m^T(j) - 2 \sum_{e} w_e \bar{\mu}_m^T(e) C_m^{-1} o_t + \mathrm{tr}(o_t o_t^T C_m^{-1}) \Bigg]$$

and thus the following are sufficient statistics for the likelihood:

$$r(e,j) = \sum_{t,m} \gamma_m(t) \bar{\mu}_m^T(e) C_m^{-1} \bar{\mu}_m(j)$$

$$b(e) = \sum_{t,m} \gamma_m(t) \bar{\mu}^T(e) C_m^{-1} o_t$$

$$c = \sum_{t,m} \gamma_m(t) \mathrm{tr}(o_t o_t^T C_m^{-1})$$

with the addition of w. The cross-correlation term, $r(e,j)$, grows with the square of the dimension of the eigenspace $\mathcal{O}(E^2)$. If we are interested in adaptation gains, then c can be safely discarded. The acoustic match, which is usually a posterior probability-weighted sum of local acoustic distances, can be summarized as cross-correlations in the probability-weighted innner product.

4.2. Fusion of segments

Define two segments of speech O_1 and O_2, with corresponding statistics $\mathcal{S}_1$ and $\mathcal{S}_2$, a rather interesting property of the statistics is that the concatenation of the segments, say O, has associated statistics $\mathcal{S}$ which can be computed as the arithmetic sum of statistics of the segments, i.e. $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$. It is equivalent to the MAP formula using conjugate priors, one segment serving as a prior to the other.

It follows from the previous derivations that the estimation of MLED eigenvalues on aribtrary concatenations of segments can be computed easily. Moreover, the estimation of the gain or decrease in likelihood given a hypothesized eigenvoice model on an arbitrary concatenation of speech segments can be done solely on the basis of the sufficient statistics. Note that MLLR has similar sufficient statistics [3]. Those familiar with MLLR will recognize the $G(i)$ and $z(i)$ matrices. They are however more cumbersome to deal with.

Additionally, since we have linear models, the likelihood is again a Gaussian and therefore attains the Cramer-Rao lower bound for the variance. It is inversely proportional to the amount of data. The squared error due to the introduction of a wrong segment is also inversely proportional to the amount of data. If a non-gaussian prior is used then the ML and the MSE (mean-squared error) differ, but the MMSE can be computed numerically by summing on points of interest.

4.3. Purity of segments based on adaptation gains

The application of the principle to our task is straightfoward. Eigenvoices can be applied successfully with extremely short

segments of speech. Thus the course of an utterance in the range of eigenvoices appears as the equivalent of many utterances for other typical adaptation algorithms. Hence, eigenvoices can work in incremental mode *within* one sentence. Unsupervised adaptation gains can be improved by rejecting incorrect transcriptions, e.g. [4]. Utterance verification techniques are applied to suspicious segments. In most utterance verification methods, the underlying assumption is that likelihood ratios form a good predictor of the correctness of a transcription, e.g. [5]. Furthermore, adaptation gains in likelihood seem to be relevant to utterance (or speaker) verification [6]. The use of sufficient statistics for speaker segmentation was explored in [3].

If we divide the utterance into small speech segments, with corresponding sufficient statistics $\mathcal{S}$ as defined in the previous section, then the leave-one-out strategy can be applied to minimize the empircally estimated expected divergence of new data. We assign high confidence to the correctness of segments that yield high adaptation gains. In practice we used segments that were one word long to estimate models, and left the rest of the utterance as cross-validation data. As with speaker segmentation, we enforce a homogenity of speech using log-probability gains. The divergence between the density estimated from one segment X on a cross-validation segment density Y is:

$$d(X,Y) = \frac{1}{2}\Big\{ E\log(2\pi) + \mathrm{tr}(R_X^{-1}R_Y) - \log|R_Y| + \\ + (w_X - w_Y)^T R_Y (w_X - w_Y)\Big\}$$

where $w_k = R_k^{-1} b_k$, $k = X, Y$. The R precision matrices and b vectors were defined above.

5. EXPERIMENTS

In this section, we describe our system and how it performed.

5.1. Conditions

For our experiments we chose the Wall Street Journal Nov92 evaluation test set. We show results on two training databases, namely WSJ0 and WSJ0+1. WSJ0, also called SI-84, consists of 7296 sentences uttered by 84 speakers. The total duration of speech amounts to about 12 hours. WSJ0+1 is also known as SI-284, and includes WSJ0 plus 200 additional speakers, for a total of about 39k sentences in 72 hours. The acoustic frontend uses 39 MFCC coefficients and sentence-based cepstral mean substraction (CMS). For SI-84, We train a total of 32000 Gaussians with diagonal covariances, pooled in 823 mixtures. For SI-284, we train 64000 Gaussians in 1404 mixtures. Thus each eigenvoice dimension consumes respectively 5 and 10 MB. The mixtures were defined using decision tree classification. We use gender-independent models. The language model (LM) for this task is the standard trigram backoff model estimated on 37M words, provided by MIT. There are about 20k words with an out-of-vocabulary rate (OOV) of about 2%.

	SI-84	SI-284
SI	13.7%	10.8%
MLLR	13.1%	10.5%
MLED	12.6%	10.1%
MLED on time segments	12.6%	10.2%
MLED w/ variable dim.	12.6%	10.1%
MLED w/ confidence	12.2%	9.8%
MLED w/ conf + LM weight	12.2%	9.7%

Table 1. Self-adaptation: WER with SI-84 and SI-284

Our recognizer, called EWAVES, is a simple lexical-tree based, word-internal context-dependent, one-pass trigram Viterbi decoder with bigram LM lookahead [7]. The test set consists of 8 speakers, none of whom are present in the training. There is an equal proportion of males and females. They read about 40-45 sentences each, summing to 333 sentences. The average length of a test sentence ranges from 5-15 seconds, with an average of 17 words per sentence. The baseline system results in a 13.7% Word Error Rate (WER) for SI-84 and 10.8% WER for SI-284.

The eigenvoices were built using standard methods as set forth in [2, 8]. We train speaker-adapted models for each speaker in the database, apply PCA to find the most important directions of inter-speaker variability, and optimize these directions (eigenvoices) with respect to the the maximum-likelihood (ML) criterion. Due to memory constraints, we limited PCA initialization to 200 speakers for SI-284. Our previous implementation of MLES proved ineffective, so we used a finer approximation. Since we have diagonal covariance matrices, for each feature dimension $d = 1, ...39$, gaussian distribution m, the ML-eigenvoices $\bar{\mu}_m^{(d)}(e)$ satisfy:

$$A_m \left[\bar{\mu}_m(1), \bar{\mu}_m(2) ... \bar{\mu}_m(E)\right]^T = z_m(d)$$

and A_m and $z_m(d)$ have components:

$$a_{kj} = \sum_q w_j^q w_k^q \sum_t \gamma_m(t), \;\; z_j = \sum_{t,q} w_j \gamma_m(t) o_d(t)$$

where w_j^q is the j-th eigenvalue of the q-th speaker. To reduce memory overhead we only store one weighted A matrix for each mixture, instead of one for each gaussian. For all experiments we used $E = 20$ eigenvoices.

5.2. Results

Results are shown in table 1. Word error rates (WERs) are reported for SI-84 and SI-284. We applied MLLR with one global matrix to get an idea of the difficulty of the task. For calibration standard MLED was also run. Jack-knifing is equivalent to baseline MLED according to WER.

We tested the assumption of stationarity as follows. We updated the estimate once every 100 ms, based on an window length of 400 ms. Surprisingly the method did not result in a change in WER, even for different values of update period and window span. We believe that the non-stationarity is exactly balanced with uncertainty due to the removal of observation data.

Then, for every utterance, we tuned the complexity of the model E. That is to say, based on the adaptation gain (and amount of training data), we forced all $w_e, e >= \hat{E}$ to be zero for some empirically determined $\hat{E}$. For all values of tuning parameters, permutation of eigenvoices, and maximum E, the system did not outperform the baseline MLED. However disappointing, it is consistent with our previous unsuccessful experiments with multigaussian prior for w (MAPED).

On the other hand, purging segments based on a ratios of adaptation gains resulted in an improvement. The false acceptance for words was about 20% and false rejection 40%. Errors in the exact transcription may not result in all wrong assignment of gaussian, and conversely a word pronounced poorly, but forced by the language model, may introduce noise in the estimation. However, intuitively, we consider one errors in assignment to be as detrimental as the added uncertainty due to the removal of two correct segments.

In our last set of experiments, we decreased the language model weight proportionaly to our confidence measure. The intuition is that in the case of poor acoustic match, we reduce the gap between first and second best hypotheses, and allow for more changes in the transcription, thereby prevent locking-in errors due to language modeling. We observed no significant improvement.

6. CONCLUSION AND FURTHER WORK

In this paper, we have shown that eigenvoices can be applied successfully to the problem of self-adaptation. We employed speaker-clustering techniques to extract homogenous, reliable statistics to fortify our estimation of speaker models. We limited the impact of corruption due to incorrect labeling by removing suspicious data. Due to the low WER of this task, the impact was bounded. Nevertheless, we discovered that the method was successful. Consequently, we plan to move to more challenging tasks such as Switchboard corpus recognition.

7. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[2] Roland Kuhn, Jean-Claude Junqua, Patrick Nguyen, and Nancy Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[3] Michiel Bacchian, "Using maximum likelihood linear regression for segment clustering and speaker identification," in *Proc. of ICSLP/Interspeech*, Beijing, China, Oct. 2000, vol. 4, pp. 536–539.

[4] T. Anastasakos and S. V. Balakrishnan, "The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers," in *Proc. of ICSLP*, Sydney, Australia, Dec. 1998, vol. 5, pp. 2203–2306.

[5] Eduardo Lleida and Richard C. Rose, "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 126–139, Mar. 2000.

[6] Michael Pitz, Frank Wessel, and Hermann Ney, "Improved MLLR speaker adaptation using confidence measures for conversational speech recognition," in *Proc. of ICSLP/Interspeech*, Beijing, China, Oct. 2000, vol. 4, pp. 548–551.

[7] Patrick Nguyen, Luca Rigazio, and Jean-Claude Junqua, "EWAVES: an efficient decoding algorithm for lexical tree based speech recognition," in *Proc. of ICSLP/Interspeech*, Beijing, China, Oct. 2000, vol. 4, pp. 286–289.

[8] Patrick Nguyen and Christian Wellekens, "Maximum likelihood Eigenspace and MLLR for speech recognition in noisy environments," in *Proc. of Eurospeech*, Sep. 1999, vol. 6, pp. 2519–2522.

[9] Jen-Tzung Chien, Jean-Claude Junqua, and Philippe Gelin, "Extraction of Reliable Transformation Parameters for Unsupervised Speaker Adaptation," in *Proc. of Eurospeech*, Budapest, Hungary, Sep. 1999, vol. 1, pp. 207–210.

[10] Toshiaki Uchibe, Shigo kuroiwa, and Norio Higuchi, "Determination of threshold for speaker verification using speaker adaptation gain in likelihood during training," in *Proc. of ICSLP/Interspeech*, Beijing, China, Oct. 2000, vol. 2, pp. 326–329.

[11] Henrik Botterweck, "Very fast adaptation for large vocabulary continuous speech recognition using eigenvoices," in *Proc. of ICSLP/Interspeech*, Beijing, China, Oct. 2000, vol. 4, pp. 354–359.

[12] Wu Chou, "Maximum A Posteriori Linear Regression with Elliptically Symmetric Matrix Variate Priors," in *Proc. of Eurospeech*, Budapest, Hungary, Sep. 1999, vol. 1, pp. 1–4.