

Reducing the Effects of Linear Channel Distortion on Continuous Speech Recognition

R. A. Bates M. Ostendorf
Boston University

March 25, 1999 — EDICS SA 1.6

Abstract

Linear channel compensation in speech recognition typically involves estimating an additive shift in the cepstral domain. This paper explores both Bayesian and maximum likelihood techniques to transform either the features or the model parameters. Experiments on the Macrophone corpus show error rate reductions over cepstral mean subtraction for short utterances.

1 Introduction

Continuous speech recognition performance is known to degrade when the speech being recognized is recorded over telephone lines, which is where most applications of speech recognition systems are currently being fielded. This performance reduction may be due to both additive noise and channel distortion from telephone handsets and lines. Since the SNR levels in existing telephone corpora are relatively high (typically greater than 20dB in Switchboard and Macrophone), this work focuses on the problem of channel distortion as the most important factor. The paper describes the modification of a hidden Markov Model (HMM) recognition system to use various techniques for linear channel estimation for finding cepstral shifts in either the feature or model space.

In the cepstral domain, the convolutional channel distortion becomes an additive vector which is constant over all frames of the speech signal. If n indicates the index of a frame of speech, then the observed speech cepstral vector is $y_n = x_n + h$, where x_n is the time-varying input speech and h is a fixed but unknown cepstral representation of the linear channel. The problem of identifying and removing h is difficult, because the signal characteristics depend on a hidden, time-varying state that is precisely what we want to recognize.

Channel compensation for speech recognition is often performed during the signal processing stage, i.e. $\hat{x}_n = y_n - \hat{h}$ where \hat{h} is the estimated channel shift. The most wide-spread method for estimating \hat{h} is to compute the mean of the cepstral vectors in an utterance [1, 2], which we will refer to as cepstral mean subtraction (CMS). CMS assumes that the average of the speech vectors will be zero and that there is enough data so that the effect of the time-varying speech characteristics can be ignored. A causal variant of CMS is RASTA [3]. Theoretically, a maximum likelihood (ML) estimate of h should give lower expected squared error than the utterance mean and thus improved recognition performance. Unfortunately, ML channel estimation has not yet shown significantly improved performance over simpler mean subtraction approaches in telephone applications [4]. However, the negative results might be explained by the following observations. First, experimental results have generally been reported on test sets where the utterances are relatively long, in which case it is more reasonable to assume that the utterance mean will converge to the channel estimate. It may be that for short utterances a more exact channel estimate will lead to improved performance. Second, the ML estimation results were based on models that are not compensated to match the channel estimation technique that will be used.

For these reasons, one goal of this work is to re-evaluate the ML technique using short utterances and training compensation. In addition, we explore improving on ML estimation by using a prior for the additive shift, which allows exploration of both feature and model space compensation techniques.

2 Channel Compensation Techniques

Since so much work has shown the benefits of cepstral mean subtraction and because it is so computationally inexpensive, it is natural to choose the cepstral mean of the utterance, \hat{h}_{CMS} , as a starting point. Equivalently, we can simply use mean-normalized cepstra and focus on estimating $\Delta\hat{h}$, the difference between the additive channel vector and the cepstral mean. Using mean-normalized cepstral vectors, $z_n = y_n - \hat{h}_{CMS}$, the estimate of the original speech signal becomes

$$\hat{x}_n = y_n - \hat{h}_{CMS} - \Delta\hat{h} = z_n - \Delta\hat{h}. \quad (1)$$

In the discussion below, \mathcal{Z} will denote as the observation sequence, which consists of CMS features $\{z_n; n = 1, \dots, N\}$.

Three types of estimates are explored here. The first two are in the feature space, as described by equation 1, and differ in terms of whether a prior on the shift is used:

$$\text{ML shift estimate: } \Delta \hat{h}_{ML} = \underset{\Delta h}{\operatorname{argmax}} p(\mathcal{Z}|\Delta h), \quad (2)$$

$$\text{MAP shift estimate: } \Delta \hat{h}_{MAP} = \underset{\Delta h}{\operatorname{argmax}} p(\mathcal{Z}|\Delta h)p(\Delta h). \quad (3)$$

In both cases, the estimate is based on the assumption that $p(z|s_n = j) \sim N(\mu_j + \Delta h, \Sigma_j)$, where s_n is the HMM state, which can be equivalently implemented in feature space when there is a single shift $\Delta \hat{h}$. The third method uses the prior in a Bayesian learning approach to transforming the model:

$$p(z|s_n = j) \sim N(\mu_j + \Delta \hat{h}_{MAP}, \Sigma_j + \Sigma_{\Delta h}) \quad (4)$$

where $\Sigma_{\Delta h}$ is a function of the number of observations and the prior covariance as shown later. The corresponding solutions for the different cases are provided below first for the case of a unimodal Gaussian observation distribution, and then the extension to mixture distributions is discussed.

The general solution for the **ML shift estimate** has been previously derived and used in both channel compensation [4, 5] and speaker normalization [6, 7]. As shown in [4, 5], an iterative solution for the ML estimate is required, since both the channel and state sequence are unknown. Fortunately, the estimate has been shown to converge after one iteration [5]. Given N observations, the general solution is

$$\Delta \hat{h}_{ML} = \left[\sum_{n=1}^N \sum_j \gamma_n(j) \Sigma_j^{-1} \right]^{-1} \sum_{n=1}^N \sum_j \gamma_n(j) \Sigma_j^{-1} (z_n - \mu_j), \quad (5)$$

where $\gamma_n(j) = p(s_n = j|\mathcal{Z})$ is the likelihood of state j at time n . The result simplifies when it is assumed that the state sequence Σ is known:

$$\Delta \hat{h}_{ML} = \left[\sum_{n=1}^N \Sigma_{s_n}^{-1} \right]^{-1} \sum_{n=1}^N \Sigma_{s_n}^{-1} (z_n - \mu_{s_n}), \quad (6)$$

using $\Delta \hat{h}_{ML} = \underset{\Delta h}{\operatorname{argmax}} p(\mathcal{Z}|\Delta h, \Sigma)$ instead of equation 2. In a paradigm of multi-pass decoding, using the best state sequence from a previous pass provides a low cost alternative to equation 5 that we will exploit in the experiments described here, both for the ML and Bayesian approaches.

If an utterance is not long enough to adequately describe the channel, using prior knowledge of the channel distribution in a **MAP estimate** may give a more reliable estimate than the ML estimate. For the MAP channel estimate, the prior distribution of the channel error is assumed to

be Gaussian $p(\Delta h) \sim N(\mu_0, \Sigma_0)$. Following the same steps as in finding the ML solution, it is easy to show that the MAP estimate for the known state sequence case is:

$$\Delta \hat{h}_{MAP} = \left(\Sigma_0^{-1} + \sum_{n=1}^N \Sigma_{s_n}^{-1} \right)^{-1} \left(\Sigma_0^{-1} \mu_0 + \sum_{n=1}^N \Sigma_{s_n}^{-1} (z_n - \mu_{s_n}) \right). \quad (7)$$

This transformation is similar to the ML estimate described above except that it includes extra terms associated with the channel prior. Asymptotically, both estimates give the same result.

Most HMM systems make use of Gaussian mixtures for the model distributions. Both the ML and the MAP channel estimate are easily extended for a model with Gaussian mixtures, as illustrated by equation 5 where the state likelihood is replaced by the mixture component likelihood.

Given a prior for the channel error vector, an alternative to MAP estimation is **Bayesian learning**, which incorporates the posterior distribution of the shift given the observations directly into the new model:

$$p(z_n | s_n, \mathcal{Z}, \mathcal{S}) = \int p(z_n | s_n, \Delta h) p(\Delta h | \mathcal{Z}, \hat{\mathcal{S}}) d\Delta h. \quad (8)$$

Assuming that both the prior and the original observation distributions are unimodal Gaussian, then the posterior distribution is Gaussian with $p(\Delta h | \mathcal{Z}, \hat{\mathcal{S}}) \sim N(\Delta \hat{h}_{MAP}, \Sigma_{\Delta h})$, where $\Delta \hat{h}_{MAP}$ is the MAP estimate above and

$$\Sigma_{\Delta h} = \left(\Sigma_0^{-1} + \sum_{n=1}^N \Sigma_{s_n}^{-1} \right)^{-1}. \quad (9)$$

The result in equation 4 follows, since equation 8 corresponds to the distribution of the sum of two independent Gaussians. Another type of model modification is stochastic matching [5]; a major difference of Bayesian learning with respect to stochastic matching, other than the use of a prior, is the assumption that the channel vector is constant.

The simple result for Bayesian learning is a result of the unimodal Gaussian distribution assumption. If the prior is a mixture Gaussian, the resulting posterior and final distributions are also mixtures with different additive mean and covariance terms. However, if the observation distribution is a Gaussian mixture, the posterior no longer has a simple Gaussian mixture form and therefore the final observation distribution does not. In order to take advantage of the power of mixture distributions in the final decoding, our solution was to estimate the posterior $p(\Delta h | \mathcal{Z}, \hat{\mathcal{S}})$ based on a unimodal observation distribution $p(z_n | s_n)$, but incorporate it with a mixture observation distribution $p(z_n | s_n, \Delta h)$ in equation 8.

3 Experiments

The experiments reported here are based on the Macrophone corpus [8], including speech that was recorded on actual phone lines, including many long distance lines. This paper focuses only on the natural numbers section of the corpus, because the utterances are typically shorter in that section and because of the usefulness of recognizing natural numbers for telephone recognition applications. The natural numbers training data includes about 15,000 utterances. Results reported here are based on an independent test set with 2000 utterances and a 530 utterance subset of short utterances that included only numbers and no disfluencies.¹ The reported results use a bigram language model based on a 514-word vocabulary (1% out-of-vocabulary rate on full test set, closed on test subset). The baseline system was developed using HTK [9], a hidden Markov model toolkit. The acoustic models used for these results were gender-independent, intra-word triphone HMMs composed of 7-mixture Gaussians. The baseline CMS system gives 15.9% and 13.1% word error rate (WER) on the full test and test-subset, respectively.

It is possible to get a better model of the telephone speech by “cleaning” the training data [10], since like the test data it comprises telephone speech. Cleaning the training data involves iteratively using ML estimation to remove channel estimates from the training data and then retraining the acoustic models. In the experiments reported here, only one iteration of cleaning with the ML channel error estimate is done, to match the test channel compensation conditions. The cleaned model is trained using four Baum-Welch iterations starting from the original 7-mixture CMS model.

Given the estimated channel vectors (actually channel *error* vectors) from the training data, it is possible to characterize the range of typical telephone channels in the form of a multivariate distribution. Statistical analyses of the channel (error) estimates support the use of a unimodal diagonal-covariance Gaussian distribution for the form of the channel prior distribution [11], which also simplifies implementation of MAP and Bayesian learning compensation techniques.

Feature transformation experiments were run using both ML and MAP channel estimates with both the original and cleaned models. The same set of modified cepstral vectors were used with both models; the only difference was in the model used in the second pass of recognition. With a baseline WER of 13.1% on the test subset, both methods reduced the WER to 12.9% with the original model and 11.0-11.1% on the cleaned model. Therefore, all further experiments used the

¹The full test set includes such expressions as “four thousand ten”, “fifteen lira”, and “seventy two acres”, as well as such utterances as “I don’t have a house number.”

cleaned models.

The recognition results for all three algorithms are shown in Table 1, together with the baseline system performance. All of the techniques implemented here demonstrate an improvement over the baseline CMS results, but only the gain on the test subset is significant. This is not unexpected, since CMS does a good job of modeling the channel for longer utterances. With the cleaned model, a 16% reduction in error rate is achieved on the test subset, which is significant at the 95% confidence level. The differences between MAP and ML are not significant, and the results do not show a significant advantage for the model-space transformation (Bayesian learning) over the feature-space transformation.

4 Discussion

In summary, this work has shown first that previous negative results on ML channel estimation relative to CMS can be explained at least in part by evaluation on long utterances. Significant gains are achieved here with more complex channel estimation techniques when used on short utterances, which is a useful result since many telephone applications would require short utterances. Iteratively cleaning the training data was an important step in achieving improved results, and the fact that previous ML work [4, 5] did not use model cleaning may also explain the earlier negative results. We have extended previous approaches to linear channel compensation by introducing Bayesian techniques for transformations in both the feature and model spaces, but there was not a significant gain from these extensions. However, the prior does not add substantially to the implementation cost for the feature space transformation, and it may be of use for very short utterances. (The model space transformation can be very expensive for cases where the acoustic model uses a large number of Gaussians.) The small gain observed for model space transformations is in contrast to that reported for stochastic matching [5]. It may be explained by the use of model cleaning in the feature-space case here and/or by the simplified one-iteration implementation of the model-space transformation used here to reduce computational costs.

Although the methodology described here shows a relative improvement over cepstral mean subtraction, there are a few straightforward extensions that will lead to even better overall results. The channel compensation results could be improved by using more iterations of channel estimation and more iterations of training data cleaning, as well as using N-best weighted estimation techniques

or word confidence weighting rather than relying on the state sequence associated with the 1-best first-pass recognition hypothesis.

Our solution to channel compensation is relatively inexpensive (assuming a multi-pass search framework), primarily because of three main simplifying assumptions: the channel is linear and constant, and the effects of noise are negligible. For wireless telephone channels, these assumptions are likely to be too restrictive. It is an open question as to whether piecewise-linear methods, as in the multiple shift approach used in [5], perform as well as more complex neural network approaches (e.g. [12]). However, if piecewise-linear methods are sufficient to address the problem, then the Bayesian approaches described here extend easily to that condition and may be more useful because of the smaller amount of data available per class when there are multiple classes.

The problem of channel compensation, particularly in the model space, can be viewed as a special case of the more general problem of adaptation. Indeed, the ML feature-space method described here is similar to techniques used for speaker adaptation. For word recognition applications where more complex adaptation techniques are used, such as ML linear regression [13], channel compensation is implicit in the overall adaptation. Thus, the simple approaches described here are mainly appropriate for short utterances where more general adaptation is not feasible. A question raised by the connection between adaptation and channel compensation is whether it is possible to separately compensate for the channel and not the speaker for problems such as speaker identification or verification. We conjecture that by having separate priors for the channel and the speaker, the Bayesian approach will help solve channel and speaker separation.

References

- [1] M. Weintraub and L. Neumeier, "Constructing Telephone Acoustic Models from a High-Quality Speech Corpus," *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1994, pp. I-85-88.
- [2] A. Anastasakos, F. Kubala, J. Makhoul, R. Schwartz, "Adaptation to New Microphones Using Tied-mixture Normalization," *Proc. ARPA Human Language Technology Workshop*, 1994, pp. 325-329.
- [3] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, October 1994, pp. 578-589.

- [4] L. Neumeyer, V. Digalakis and M. Weintraub, "Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus," *IEEE Transactions on Speech and Audio Processing*, October 1994, pp. 590-597.
- [5] A. Sankar and C. -H. Lee, "A Maximum-likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, May 1996, pp. 190-202.
- [6] S. J. Cox and J. S. Bridle, "Unsupervised Speaker Adaptation by Probabilistic Spectrum Fitting," *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1989, pp. I-294-297.
- [7] Y. Zhao, "An Acoustic-Phonetic-Based Speaker Adaptation Technique for Improving Speaker-Independent Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp 380-394, July 1994.
- [8] J. Bernstein, K. Taussig and J. Godfrey, "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project," *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1994, pp. I-81-84.
- [9] S. Young, J. Jansen, J.J. Odell, D. Ollason, P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, England, 1995.
- [10] P. Moreno, M. Siegler, U. Jain, R. Stern, "Continuous Recognition of Large Vocabulary Telephone Quality Speech," *Proc. ARPA Spoken Language Systems Technology Workshop*, January 1995, pp. 70-74.
- [11] R. A. Bates, "Reducing the Effects of Linear Channel Distortion on Continuous Speech Recognition," Boston University College of Engineering, Master's Thesis, 1996.
- [12] A. C. Surendran, C.-H. Lee and M. Rahim, "Maximum-likelihood Stochastic Matching Approach to Non-linear Equalization for Robust Speech Recognition," *Proc. International Conference on Spoken Language Processing*, October 1996, pp. 1836-1839.
- [13] C. J. Leggetter and P. Woodland, "Speaker Adaptation Using Maximum Likelihood Linear Regression," *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, 1995.

Table 1: Transformation results with CMS baseline, using a cleaned acoustic model and a bigram language model (percentage word error).

Test Set	CMS	\hat{h}_{ML}	\hat{h}_{MAP}	Bayesian Learning
Subset	13.1	11.0	11.1	11.0
Full	15.9	15.6	15.6	15.5

List of Tables

1	<i>Transformation results with CMS baseline, using a cleaned acoustic model and a bigram language model (percentage word error).</i>	9
---	--	---