# AUTOMATIC TRANSCRIPTION OF VOICEMAIL AT AT&T

*Michiel Bacchiani*

AT&T Labs-Research, 180 Park Ave., Florham Park, NJ 07932, USA
michiel@research.att.com

**ABSTRACT**

This paper reports on the automatic transcription accuracy of voicemail messages. It shows that vocal tract length normalization and adaptation using linear transformations, proven to improve accuracy on the Switchboard task, provide similar accuracy improvements on this task. Direct application of the normalization techniques is complicated by the fragmentation of the data. However, unsupervised clustering was found to be effective in ensuring robust estimation of normalization parameters. Variance adaptation resulted in larger accuracy improvements than adaptation of only mean parameters, probably due to a large variability in channel conditions. The use of semi-tied covariances provides additional gains over using speaker and channel normalization. The combined gain of using various compensation techniques improves the system word error rate from 34.9% for the baseline system to 28.7%.

## 1. INTRODUCTION

In recent years, in light of the Switchboard evaluations [9], several compensation algorithms have been shown to improve the accuracy of automatic transcription of spontaneous speech. The improvements can be attributed to the partial success of these algorithms to compensate for speaker/channel variations and invalid modeling assumptions. The most widely used techniques that have shown performance improvements across different sites when implemented independently are Vocal Tract Length Normalization (VTLN) [1], adaptation using Maximum Likelihood Linear Regression (MLLR) [3] and Semi-Tied Covariances (STC) [6]. The VTLN technique normalizes for speaker variability by applying a non-linear transformation to the acoustic features. The MLLR technique compensates for speaker and/or channel variability by a speaker-dependent linear transformation of the model mean parameters. The STCs decorrelate the elements of the acoustic features by a linear transformation to better match the widely used diagonal covariances. The set of these techniques will be referred to in the rest of this paper as compensation algorithms. The techniques that attempt to normalize for speaker dependent characteristics (VTLN and adaptation) will be referred to as normalization techniques.

The focus of this work is on how the compensation algorithms, proven to be effective on the Switchboard task, can be applied to a voicemail transcription task. This task has many similarities to the Switchboard task (spontaneous speech recorded through a telephone channel) but also has several characteristics that differ. First, unlike the Switchboard corpus, the data consists of a large number of short messages. The data fragmentation and the assumption of a lack of supervisory information about speaker identity and channel conditions make robust estimation of normalization parameters difficult. Second, the voicemail transcription task used in these experiments exhibits a large variability in terms of channel conditions (use of cellular phones for example). Therefore, the previously developed normalization algorithms require an extension to ensure robust parameter estimation. In addition, variance adaptation might be more important given the large channel variations [8].

Three key sets of experimental results are reported in this paper. To ensure robustness of the normalization parameter estimation, a clustering approach was investigated. Section 5 describes two clustering algorithms and their impact on the transcription accuracy. In section 6 two linear transformation-based adaptation algorithms are compared, one adapting only means, the other adapting both means and variances. Section 7 describes how the use of STCs affects system performance. Since using linear mean+variance adaptation transformations in training allows for feature decorrelation as well, the comparison of systems with and without STCs is required to separate the benefits from variance adaptation and feature decorrelation.

Before describing the key experimental results, the voicemail corpus is described in section 2. Then, in section 3, the system training algorithm and performance of the Gender Independent (GI) baseline system and a Gender Dependent (GD) system are reported. Section 4 describes the details of the compensation algorithms used.

## 2. VOICEMAIL CORPUS

The transcription experiments were conducted on a 100 hour corpus of voicemail messages collected from the voicemail boxes of 140 employees at AT&T. The corpus contains approximately 10,000 messages from approximately 2500 speakers. The messages were manually transcribed and labeled for channel condition, speaker gender, speaker identity and whether or not the message was from a native speaker. About 90% of the messages were recorded from regular handsets, the rest from cellular and speakerphones. The corpus is approximately gender balanced and approximately 12% of the messages are from non-native speakers. The mean duration of the messages is 36.4 seconds, the median is 30.0 seconds. The recordings were digitized at a sampling rate of 8kHz and encoded as 8-bit $\mu$-law samples.

## 3. BASELINE AND GENDER DEPENDENT SYSTEMS

To evaluate the transcription accuracy, here and in all other results reported in this paper, the corpus was partitioned randomly into a 60 hour training set (700k words) and a 40 hour test set (388k words). The transcriptions of the training messages were used to construct a trigram language model. A 14k dictionary was con-

| System | Word Error Rate (%) |
|---|---|
| GI | 34.7 |
| GD | 33.3 |

**Table 1**. Baseline system error rate on the 40 hour test set.

structed using the AT&T Labs NextGen Text To Speech system for all unique words observed in the training set. The dictionary used 42 phonemic sub-word units, 5 noise units and 1 silence unit.

The acoustic feature vectors consisted of the first 12 FFT-based, Mel-warped cepstral coefficients, an energy coefficient and their first and second order time-derivatives. On a per message basis, the mean of the cepstral parameters was subtracted and the parameters were scaled to unit variance.

All acoustic models used 3-state left-to-right triphone HMMs. The states of the triphone HMMs were tied using likelihood-based decision-tree clustering of full covariance Gaussian distributions. The output distributions of the tied states were modeled using 12-component Gaussian mixture distributions with diagonal covariances. The parameters of the mixture distributions were obtained by a hybrid Viterbi and Expectation Maximization (EM) training algorithm which used EM training only within word-boundaries. The word-boundaries were estimated by Viterbi alignments. The complexity of the tied-state output distributions were incrementally increased. The $N+1$ component mixture distributions were initialized by perturbation of the most heavily weighted mixture components in the $N$ component mixtures. EM training was then used to estimate the parameters of the mixtures. Viterbi alignments of word-boundaries were only performed using some of the intermediate stage systems. All models were trained by increasing complexity with increments of 1 mixture component up to 8 mixture components and then used increments of 2 mixture components to obtain the final 12-component distributions. At every stage, 4 iterations of EM training were run to estimate model parameters. Viterbi alignments were performed after estimating the 6, 8 and 10 component mixture distributions.

Using the described training algorithm, a 8016 tied-state GI model was trained. Using the supervisory gender information, a GD model was built with 4016 tied-states each. The error rates of these models on the 40 hour test set are given in Table 1. Except for the first pass that used the GI system, all experimental results are based on rescoring the lattices generated by this first pass. The GD system was used to determine gender based on likelihood for all messages in the test set and for all further experimentation. Retaining the most likely transcripts, the GD transcription accuracy reported in Table 1 was obtained.

## 4. COMPENSATION ALGORITHMS

Details of the compensation algorithms used in the reported experiments are described here. The VTLN algorithm is described in section 4.1. The adaptation algorithms, based on Maximum Likelihood (ML) estimation of linear transformations, are described in sections 4.2 and 4.3 for the means only and means+variances algorithms respectively. The details of STC algorithm are described in section 4.4.

### 4.1. Vocal Tract Length Normalization

The vocal tract length normalization algorithm used a piecewise linear frequency warping implemented similar to [2]. Denoting frequency with respect to the Nyquist frequency, the frequency warping is linear with slope $\alpha$ from 0 to 0.8 and linear with slope $\frac{1}{\alpha}$ from 0.8 to 1.0. To obtain amplitude estimates at equidistant points in the warped frequency domain, the FFT values were interpolated with a cubic spline. The warp selection was based on likelihood of voiced phones in a Viterbi alignment using either the GD model described in section 3 or a VTLN trained model. The alignments were either against the last available hypothesized transcript (in test) or the reference transcript (in training). The warps were constrained to be between 0.9 and 1.1 with a step size of 0.02.

### 4.2. Mean Adaptation

The algorithm used to adapt means was an implementation of the MLLR algorithm described in [3]. All mean adaptation experiments used one full transformation matrix plus offset applied to all mixture components. Experiments where MLLR adaptation was used in training and testing followed an implementation of the Speaker Adaptive Training (SAT) algorithm described in [4].

### 4.3. Constrained Mean and Variance Adaptation

The algorithm used to allow adaptation of both means and variances was the Constrained Model-space (CM) adaptation algorithm described in [5]. The implementation of the iterative optimization procedure directly followed the described algorithm. In all experiments, 10 optimization iterations were run to obtain the final transformation estimates. All experiments used a full transformation matrix plus offset. In some experiments, multiple regression classes were used, each using a full transformation matrix plus offset. The regression classes were hand designed and fixed throughout the experiments. In the experiments using two regression classes, the classes were silence+noises and speech. In the experiments using five regression classes, there were separate transformations for noises+silence, vowels+semivowels+glides, nasals, stops and affricates+fricatives. In the experiments where CM adaptation was applied in both training and test, the system will be referred to as CM-SAT.

### 4.4. Feature Decorrelation

The algorithm used to decorrelate the elements of the acoustic feature vectors was the semi-tied covariance algorithm described in [6]. However, the optimization algorithm for estimating the semi-tied transformation and corresponding model differed slightly from that algorithm. Here, starting with the fully trained VTLN model, the following iterative algorithm was executed:

1. Using the last estimate of the semi-tied transformation (identity for the first iteration) and last model estimate, compute posterior probabilities of occupying each mixture component at each time and collect the statistics for estimating the semi-tied transformation.
2. Estimate the semi-tied transformation based on the statistics collected in step 1.
3. Using the same posterior probabilities as found in step 1, re-estimate the model parameters using the semi-tied transformation estimate from step 2.
4. Complete the model update by running 3 iterations of EM training using the last semi-tied transformation estimate.
5. If another training iteration, go to step 1.

## 5. MESSAGE CLUSTERING

The use of normalization techniques (VTLN and adaptation) in the voicemail transcription task is complicated by the fragmented nature of the corpus. Direct estimation of normalization parameters based on the data available in a single message can lead to large estimation errors due to insufficient data. Because multiple messages are available for many speakers, estimation error could be avoided if these messages were pooled together and share a common set of normalization parameters. To implement this, clustering approaches were investigated. In cases where normalization is only applied on the test data, only the test data needed to be clustered. As availability of supervisory information about speaker and/or channel conditions could not be assumed in such a scenario, an unsupervised clustering algorithm was required. If the normalization techniques were also used in training (training a VTLN or SAT model), clustering of the training data was also required. There, availability of supervisory information about speaker and channel could be assumed and could be used in clustering. However, since many speakers have only very little data in the whole corpus and since in many cases the channel conditions change from message to message for a given speaker, the optimality of a clustering configuration remains unclear even given the supervisory information. Therefore, in the experiments using normalization in training, unsupervised clustering was applied to the training data as well.

Two clustering approaches were investigated that were compared in more detail in [7]. The first used Text Independent Gaussian Mixture Models (TIGMMs) to represent messages and used an agglomerative clustering approach with a likelihood based distance metric. The models were estimated on the speech frames of the messages only. To distinguish speech from silence and noises, the final Viterbi alignments of the GD model training described in section 3 were used for the training data. For the test data, the final alignment of the GD model hypothesis was used. The features used for the TIGMMs were 12 dimensional linear predictive coding derived cepstral features and their first order time derivatives. The implementation used 64-component, diagonal covariance mixture models, estimated using a training algorithm similar to the one described in section 3, except that no Viterbi alignments were used and complexity increments were in steps of 4 mixture components. In addition, if the occupancy count of a mixture component that was to be split fell below 100 frames, the covariances of the newly formed mixture components were tied. After estimating the TIGMMs for all messages, agglomerative clustering was used to ensure data pools of at least 40 seconds of data. The symmetric distance metric between messages used in clustering was

$$D(i,j) = -\sum_{n=1}^{64} \mathcal{L}(\mu_n^{(j)} \mid \mathcal{M}) - \sum_{m=1}^{64} \mathcal{L}(\mu_m^{(i)} \mid \mathcal{N}) \qquad (1)$$

where $D(i,j)$ denotes the distance between messages $i$ and $j$, $\mathcal{L}(.)$ denotes the log-likelihood function, $\mu_m^{(i)}$ denotes the $m$-th mean of the mixture model of message $i$, $\mu_n^{(j)}$ denotes the $n$-th mean of the mixture model of message $j$, $\mathcal{M}$ denotes the mixture model of message $i$ and $\mathcal{N}$ denotes the mixture model of message $j$.

The second clustering approach was the MLLR-based algorithm described in detail in [7]. This algorithm uses the MLLR adaptation statistics to directly optimize the MLLR adaptation likelihood of the cluster data. In contrast to the TIGMM approach, this clustering approach is consistent as it optimizes the same objective used in MLLR adaptation. In addition, this approach is efficient as it uses the adaptation statistics based on the recognition model and does not require the construction of an external model. As in the case of the TIGMM clustering approach, the algorithm was used to cluster the messages agglomeratively into clusters of at least 40 seconds of speech data.

| Model | Test Normalization | Word Error Rate (%) | |
|---|---|---|---|
| | | TIGMM | MLLR |
| GD | - | 33.3 | 33.3 |
| GD | VTLN | 32.4 | 32.3 |
| VTLN | VTLN | 32.3 | 32.0 |
| VTLN | VTLN+MLLR | 31.2 | 30.9 |
| VTLN+SAT | VTLN+MLLR | 30.9 | 30.4 |

**Table 2**. Normalized system performance using either TIGMM or MLLR based message clustering.

| Model | Adaptation Type | Word Error Rate (%) |
|---|---|---|
| VTLN | MLLR | 30.9 |
| VTLN+SAT | MLLR | 30.4 |
| VTLN | CM | 30.7 |
| VTLN+CM-SAT | CM | 29.7 |

**Table 3**. System performance using mean only adaptation or constrained mean+variance adaptation applied in either test only or training and test.

The transcription accuracies using the two clustering approaches are given in Table 2. Using either clustering approach, the transcription accuracy improves by using the normalization techniques. Use of the adaptation likelihood-based clustering consistently improves performance over the TIGMM-based clustering approach and gives a 0.5% accuracy improvement for the system that uses both VTLN and adaptation in both training and test. Due to this observation, MLLR-based clustering was used to define the message clusters that share normalization parameters in all subsequent experiments.

## 6. ADAPTATION

Previous work on variance adaptation has shown that this is particularly beneficial for tasks that exhibit large variation in channel characteristics [8]. As this voicemail corpus exhibits such variations, the linear transformation-based adaptation technique that adapts only model means (MLLR) was compared with an approach that adapts both means and variances using a constrained linear transformation (CM). Table 3 shows the performance of these adaptation approaches using adaptation in test only or in both training and test. The system that used adaptation in training using the CM algorithm was trained using two iterations of CM-SAT. The performance of the system obtained by one CM-SAT pass was 29.9; 0.2% worse. Using a second pass of SAT of the MLLR-based system did not show any additional improvement in transcription accuracy. It can be observed from Table 3 that the CM adaptation algorithm performs slightly better than MLLR if used in test alone (30.7 vs. 30.9) and somewhat more if used in both training and test (29.7 vs. 30.4).

Although the CM adapted system has the ability to adapt both means and variances, the estimated linear transformation is constrained as it applies to both means and variances. Hence, one option to further improve modeling accuracy is to cascade CM adaptation with MLLR adaptation. Another option for improvement of modeling accuracy is to use multiple regression classes. The per-

| Adaptation Approach | Word Error Rate (%) |
|---|---|
| CM+CM2 | 29.5 |
| CM+CM2+CM5 | 29.4 |
| CM+MLLR | 29.3 |

**Table 4**. System performance using a cascade of adaptation passes. A + indicates iterative adaptation and the numerical suffixes indicate the number of used regression classes.

| Model | Adaptation Approach | Word Error Rate (%) |
|---|---|---|
| VTLN+CM-SAT | CM | 29.7 |
| VTLN+CM-SAT | CM+MLLR | 29.3 |
| VTLN+STC+CM-SAT | CM | 29.3 |
| VTLN+STC+CM-SAT | CM+MLLR | 28.7 |

**Table 5**. Adapted performance of systems with and without using semi-tied covariances.

formance of these different options starting from the VTLN+CM-SAT model are given in Table 4. The progressively more complex adapted models were obtained by iterative decoding using the hypotheses of the last recognition pass to estimate the transformations of the next pass. It can be observed that cascading CM and MLLR provides slightly better performance than iterative adaptation using CM with multiple regression classes. Further investigation showed that cascading the 5 regression class CM adaptation with MLLR adaptation did not show any additional gains. Furthermore, using 5 regression classes in CM-SAT performed slightly better after one training pass (29.8 vs. 29.9) but slightly worse after a second iteration (29.8 vs. 29.7). Finally, the use of block-diagonal transformation matrices performed slightly worse than using a full transformation matrix (29.9 vs. 29.8).

## 7. SEMI-TIED COVARIANCES

The improved performance of the CM adaptation in comparison to the MLLR-based adaptation approach is potentially due to the fact that CM-SAT can better compensate for speaker and channel variations (due to variance adaptation) but it could also be due to decorrelation of the acoustic features. Both the CM-SAT algorithm and the STC algorithm estimate linear transformations applied to the acoustic features and then estimate model parameters based on the linearly transformed features. The STC algorithm finds a linear transformation applied independently of speaker identity whereas the CM-SAT algorithm estimates speaker-dependent transforms. However, since both transforms are linear, the CM-SAT algorithm has a sufficient number of free parameters to learn both simultaneously. The difference using STCs followed by CM-SAT versus using CM-SAT alone is the starting model and initial transform estimate. To compare the impact of this, two systems, one with and one without STCs were compared. Starting from the GD VTLN models, using the iterative algorithm described in section 4.4, two STC systems was trained, one using a single STC transformation, the other using seven transformations for different hand-designed regression classes. The best performance was obtained using the single transform STC system after seven iterations of training which provided a 1.2% accuracy improvement over the GD VTLN model. The performance of this system with adaptation and the system without STCs but with adaptation are given in Table 5. It can be observed that the gains from CM-SAT and STCs are not additive, with the 1.2% gain of using STCs on the VTLN model reduced to a 0.6% gain after CM-SAT. However, the use of STCs before CM-SAT did provide additional gains, indicating the importance of the seed model that the CM-SAT is initialized with.

## 8. CONCLUSIONS

The experiments reported in this paper show that gains from VTLN, linear transformation-based adaptation and STCs are similar to those obtained on the Switchboard task [9]. Clustering was effectively used to ensure sufficient data for the estimation of normalization parameters. The transcription accuracy was dependent on the type of clustering algorithm. CM adaptation provided larger accuracy gains than MLLR adaptation especially if used in both training and test. Use of a cascade of CM and MLLR adaptation provided additional accuracy improvements. Further accuracy improvements could be obtained by the use of STCs. The STC and CM adaptation gains were not additive, indicating that use of CM adaptation in training partially succeeds in decorrelating the acoustic features. Assuming the STC gain (1.2%) is additive to the MLLR-based adaptation gain (1.6%), the combined gain would still be less than that of the joint gain of using STCs and CM adaptation (3.3%) showing the advantage of using variance adaptation in this task.

## 9. REFERENCES


[1] T. Kamm, G. Andreou and J. Cohen, "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability", In *Proc. of the 15th Annual Speech Research Symposium*, pp. 161-167, CLSP, Johns Hopkins University, Baltimore, MD, June 1995.

[2] S. Wegmann, D McAllaster, J. Orloff and B. Peskin, "Speaker Normalization on Conversational Telephone Speech", In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 339-341, 1996

[3] C. J. Legetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, pp. 171-185, 1995.

[4] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A compact model for speaker-adaptive training," In *Proceedings of the International Conference on Spoken Language Processing*, pp. 1137-1140, 1996.

[5] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, 12, pp. 75-98, 1998.

[6] M. J. F. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 7, No. 3, 1999.

[7] M. Bacchiani, "Using Maximum Likelihood Linear Regression for Segment Clustering and Speaker Identification," In *Proceedings of the International Conference on Spoken Language Processing*, Vol. 4, pp. 536-539, 2000.

[8] P. C. Woodland, M. J. F. Gales and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 65-68, 1996.

[9] Proceedings of the Speech Transcription Workshop, University of Maryland, May 16-19, 2000.