



Your Mobile Virtual Assistant Just Got Smarter!

Mazin Gilbert, Iker Arizmendi, Enrico Bocchieri, Diamantino Caseiro, Vincent Goffin, Andrej Ljolje, Mike Phillips¹, Chao Wang¹, Jay Wilpon

AT&T Labs-Research, Florham Park, NJ 07932

¹ Vlingo, Cambridge, MA 02138

Abstract

A Mobile Virtual Assistant (MVA) is a communication agent that recognizes and understands free speech, and performs actions such as retrieving information and completing transactions. One essential characteristic of MVAs is their ability to learn and adapt without supervision. This paper describes our ongoing research in developing more intelligent MVAs that recognize and understand very large vocabulary speech input across a variety of tasks. In particular, we present our architecture for unsupervised acoustic and language model adaptation. Experimental results show that unsupervised acoustic model learning approaches the performance of supervised learning when adapting on 40-50 device-specific utterances. Unsupervised language model learning results in an 8% absolute drop in word error rate.

1. Introduction

HAL 9000 from “2001: A Space Odyssey” and the 1987 Apple “Knowledge Navigator” are conceptual examples of how people can naturally interact with personal virtual agents - communication agents that know who you are, can recognize and understand free speech, and perform actions such as retrieving information or completing transactions.

The proliferation of mobile and hand-held devices along with advances in multimodal and multimedia technologies are giving birth to a new wave of communication agents that enable users to quickly and more naturally perform many tasks including finding music, videos, and business listings, surfing the web, sending an SMS, reserving restaurant tables, or interacting with social media websites through voice¹. One example of such communication agents is the Vlingo mobile application. Using voice input/output, users can send and listen to SMS and emails, interact with social media sites like Facebook, Twitter and Foursquare, find, call and navigate to businesses or locations, search the web and dial personal contacts.

As the number of smart phones and emerging devices continues to grow, the demand for communication agents will continue to rise. Speech and multimodal interfaces have and will continue to play a central role in developing advanced MVAs. One essential characteristic of any of these intelligent virtual assistants is the ability to learn and adapt without supervision. Unsupervised learning poses several technical challenges. First, variations in the acoustic conditions of the speech signal due to different mobile platforms, microphone characteristics, encoding methods, speaker accents/dialects, hands free distances, and background environment result in noisy speech transcription with unreliable confidence scoring. Second, having multiple speakers per device poses difficulties

in performing user adaptation. Finally, the limited amount of audio data available per device especially those with poor speech recognition performance provide noisy statistical estimates for model updates.

Over the past two decades there have been significant research in both supervised and unsupervised acoustic and language model adaptation for improving speech recognition performance. In [1], for example, we showed that location information helps to improve accuracy of speech recognition by reducing the search space. In this paper, we will highlight our adaptive learning research for developing advanced MVAs. In particular, we will present our acoustic and language modeling algorithms for rapid speech adaptation. These algorithms form the foundation for enabling accurate and robust speech understanding and intent modeling, as well as for providing a personalized experience for retrieval of information and fulfillment of transactions.

The outline of this paper is as follows. Section 2 describes the broad technologies behind MVAs. Section 3 provides details of the adaptive learning techniques for acoustic and language modeling used in the AT&T WATSON™ speech engine [2]. Section 4 presents results for supervised and unsupervised adaptation at the mobile device level. A brief summary is presented in Section 5.

2. MVA Technologies

There are three major technology components of most communication agents. The first is multimodal processing. This is the input side that includes processing and integrating heterogeneous signals such as speech, text, gesture, video, and location, and converting them into actionable requests, such as dictating an SMS or searching for a song by title. The second component is information retrieval and fulfillment. This is the media processing which involves searching through data warehouses, such as the web or local knowledge databases, to retrieve information or perform transactions such as ordering a movie ticket. The third component is multimedia processing. This is the output side that involves visualization, display and ranking of information that may include text, speech or other media signals.

In this paper, we present the multimodal-processing component for conducting speech recognition that is needed to develop advanced MVAs. In the next section, we present the algorithms used for unsupervised learning of acoustic and language modeling.

3. Adaptive Learning in Speech Recognition

A schematic diagram of the adaptive learning system for speech recognition is shown in Figure 1. The basic technology components include a feature extractor and a speech decoder. The decoder is based on a finite state transducer that searches through a composed network of acoustic, language and lexical

¹ Examples of such applications include www.vlingo.com, www.ypmobile.com, [goog 411](http://goog411.com), and www.siri.com

models. Following recognition, a set of data including audio, n-best hypotheses, lattices, confidence scores, as well as application and device specific information are used for adaptive learning. For each mobile device registered by the application platform, the adaptive learning system updates all corresponding acoustic, lexical and language models.

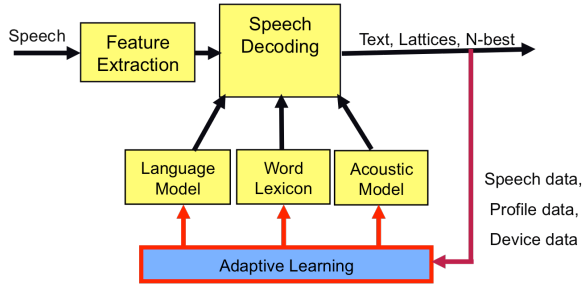


Figure 1: Schematic diagram of an adaptive speech recognition component of a MVA.

3.1 Language Model Adaptation

Successful language model adaptation with limited in-domain data is challenging. One of the reasons is that the parameters of the n-gram language models have little structure. Given the nature of our application, we use hierarchical language models (HLM) [3,4] which enable us to independently build language models for sub tasks or constituents and combine them using suitable carrier phrases. Since the combination is done dynamically, this structure is very convenient for adaptation, as different components can be quickly assembled at runtime depending on meta-data, such as the device id, device maker, or its location.

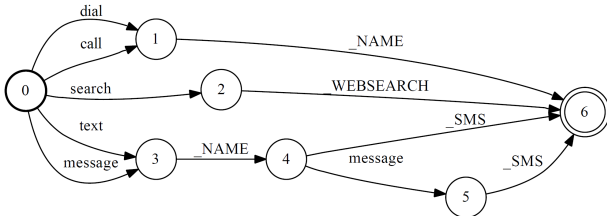


Figure 2. Example of HLM top level.

Figure 2 exemplifies the top level of a typical HLM. Tokens in an HLM can be regular words or tags referring to component language models, in this example, `_NAME`, `_SMS` and `_WEBSEARCH`. The top-level models carrier phrases, and it can be either hand crafted grammars or an n-gram model trained from data. In this case, a transcription corpus is parsed² to identify the constituents, which in turn are replaced by tags. During training, tags are treated like any other word in the vocabulary, but during recognition, they are dynamically replaced by their respective constituent sub-language models.

The constituent language models can be hand crafted or automatically trained. As before, training relies on parsed transcriptions from which the relevant constituent text is extracted to form the transcription corpus.

Since there may be few data to build accurate models for some of the constituents, we use linear interpolation and Maximum A-Posteriori (MAP), as shown in Equation 1, to

² Parsing can be bootstrapped by using the HLM as a parser.

adapt models built with lower quality transcription data [5]. We compute the probability of word w given history h from the smoothed probabilities P_i of models built with different data sets, and interpolation weights λ :

$$p(w|h) = \sum_i \lambda_i p_i(w|h) \quad (1)$$

$$p(w|h) = \frac{\sum_i \lambda_i p_i(h) p_i(w|h)}{\sum_i \lambda_i p_i(h)}$$

Lower quality data may include out of domain data and automatically recognized speech, which is either accepted by the user or has high confidence (unsupervised adaptation). For example, to build the SMS language model component we combine transcribed spoken SMS, recognized SMS and data harvested from social media web sites. Using aggregated or clustered device specific transcriptions, we adapt the HLM to classes of devices based on device specific information [1].

3.2 Acoustic Model Adaptation

The frontend of the AT&T WATSON™ recognizer produces 60 dimension feature vectors, at a frame rate of 100 per second [2]. A linear transformation using Heteroscedastic Linear Discriminant Analysis with Maximum Likelihood Linear Transform (HLDA-MLLT) is applied to every 11 consecutive frames. A frame includes mel-cepstral frequencies, following cepstral subtraction, and a normalized energy coefficient. Acoustic modeling is done using context-dependent hidden Markov models (HMM) with tied states. State tying is performed by decision trees [6].

The HMM state output densities, essentially device independent (DI), are Gaussian mixture models (GMM) with diagonal covariances:

$$p_{DI}(\mathbf{x} | s) = \sum_{i=1}^{N_s} w_{i,s,DI} \mathbf{N}(\mathbf{x}, \mu_{i,s,DI}, \Sigma_{i,s,DI}) \quad (2)$$

where s, i are the state and Gaussian component index, respectively.

Intuitively, each mobile device is likely to be used by one or very few users in a characteristic mixture of acoustic environments, while commuting, working, shopping, etc. This motivates the investigation of device-specific (DS) acoustic models, namely models that are estimated and tested on speech data from a specific device (individual devices may be identified using a unique identifier such as the IMEI, which is the *International Mobile Equipment Identity*). In practice, we are interested in modeling relatively small sets of speech data containing as few as one hundred DS sentences. Hence, to limit the DS model parameter variance, the estimation methods use constraints in the form of priors with a few free parameters. We review below the constrained estimation/adaptation algorithms that are adopted in this study. These algorithms can be used in supervised or unsupervised mode, by supplying either the reference transcription of the speech data or a word lattice output from the speech recognizer.

3.2.1 MAP Adaptation

In adaptation based on a MAP criterion, the parameters of the prior DI model (1) are updated by interpolation with the statistics of the DS data [7]. For adaptation of the Gaussian means, the equation holds:

$$\mu_{i,s,DS} = \frac{\tau_{i,s} \mu_{i,s,DI} + c_{i,s} \bar{x}_{i,s}}{\tau_{i,s} + c_{i,s}}$$

where $\tau_{i,s}$ is the prior weight and $c_{i,s} \bar{x}_{i,s}$ are the count and the first order statistics of the DS data. Similar expressions hold for the second order statistics and mixture component weights. MAP adaptation has been extremely useful in the adaptation of an acoustic model to a specific application [1][8], and we are experimenting with its use to enhance the accuracy of sub-tasks within an application. MAP variants have been devised to improve its efficiency with relatively small data sets, e.g. structural MAP [10]. However, it still requires more data than what is available to us for many individual devices.

3.2.2 MLLR Adaptation

Maximum likelihood linear regression (MLLR) of the Gaussian means is very suitable to our task because it requires the estimation of one or few affine transforms [9]. The HMM Gaussian means become:

$$\mu_{i,s,DS} = A_{g(s),DS} \mu_{i,s,DI} + \mathbf{b}_{g(s),DS} \quad (3)$$

where $g(s)$ is the mapping from HMM state to regression class, typically defined by phonetic categories. The number of free parameters can be further reduced by imposing a structure (forcing to zero sum) of the transform parameters. However, MLLR requires either storing large acoustic models for millions of devices, or transforming the Gaussian means at run-time for the computation of the likelihood. Therefore, we prefer the use of a different technique, known as constrained model adaptation (CMA).

3.2.3 Constrained Model Adaptation

With CMA[11], the adaptation of the HMM state likelihood computations to each unique device is performed in the feature domain by linear transformation $A_{g(s),DS}$:

$$p_{DS}(\mathbf{x} | s) = \sum_{l=1}^{N_s} w_{i,s,DI} \mathbf{N}(A_{g(s),DS} \mathbf{x}, \mu_{i,s,DI}, \Sigma_{i,s,DI})$$

These transformations are stored for every device in the database, and efficiently applied at run-time to the feature vectors. Both CMA and MLLR can be further improved by substituting the DI parameters in Equation (2) with values iteratively estimated (in the off-line training phase) on the transformed feature vectors $A_{g(s),DS} \mathbf{x}$. This approach is known as speaker adaptive training (SAT) [11].

4. Experimental Results

Device-specific information, such as device identity and type, is essential for improving speech recognition performance and enabling advanced MVAs. We conducted several experiments to evaluate the impact of adaptive learning on speech recognition performance. A corpus of speech data has been extracted from a set of mobile phones including, iPhone, Android, Symbian and Blackberry devices. Subjects were accessing their MVA to perform various tasks including web search, voice dialing, email and SMS dictation, and social media update.

4.1 Acoustic Model Adaptation

To evaluate the performance of acoustic model adaptation, a collection of 200 utterances per device for 95 devices has been collected and transcribed. The first 150 utterances were made available for adaptive learning, while the last 50 utterances formed the test set. The learning was performed by using CMA with either human transcription (supervised) or transcription generation by AT&T Watson™ with associated confidence scores (unsupervised). Confidence scores were based on posterior probabilities of word confusion networks [12]. Two sets of experiments were conducted that compare the baseline performance with DS adapted performance:

- Supervised adaptation: adapting the acoustic model for each device (with manual transcription) using randomly selected 5 to 150 utterances, or utterances that are ranked by their confidence scores.
- Unsupervised adaptation: adapting the acoustic model for each device (with automated transcription) using randomly selected 5 to 150 utterances, and utterances that are ranked by their confidence scores.

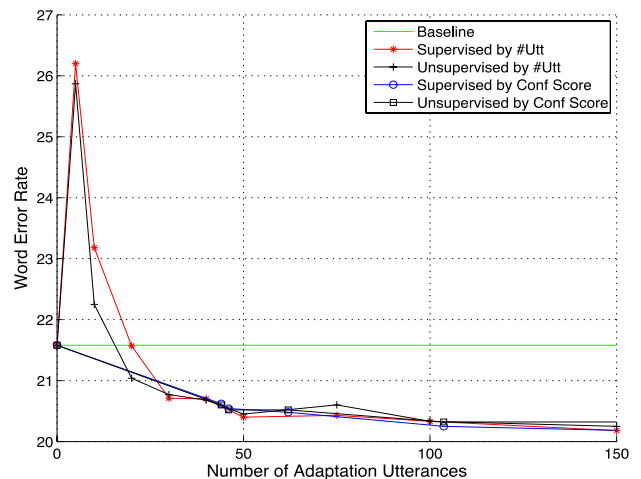


Figure 3: Word error rate for supervised and unsupervised acoustic model adaptation.

The recognition performance is shown in Figure 3. The baseline word error rate is 21.6% when evaluated on the 50 test utterances for each of the 95 devices. Several observations worth pointing out: (a) supervised learning with less than 20 utterances led to worse performance than the baseline system; (b) as the amount of data increased beyond 20, there was a reduction in word error rate for both supervised and unsupervised learning; and (c) supervised learning with utterances selected randomly or using confidence scores performed nearly the same. Note the difference observed within 20 utterances was attributed to having large data with high confidence scores. Lastly, with about 40-50 utterances, both supervised and unsupervised adaptation converges to similar word error rate. The final error rate with 150 utterances was 20.3%.

We performed a further study to understand the extent of the improvement across each device. Figure 4 shows the percentage of devices experiencing a drop (ASR +0%) or increase (ASR -0%) in word error rate as a function of the number of utterances used during unsupervised learning. For example, ASR +10% represents the percentage of devices that experienced a drop of more than 10% in word error rate for a given number of utterances. The results show that over 65% for ASR +0%, whereas nearly 25% of devices for ASR -0%.

50% of devices had a gain in accuracy of more than 5% versus only a few more than 10% of devices had a loss in performance of at least 5% (ASR+5% and ASR-5%). Furthermore, almost 30% of devices had a gain of over 10%, while less than 10% had a loss of at least 10%. These results demonstrate that unsupervised adaptation had a positive impact on performance on a large percentage of devices.

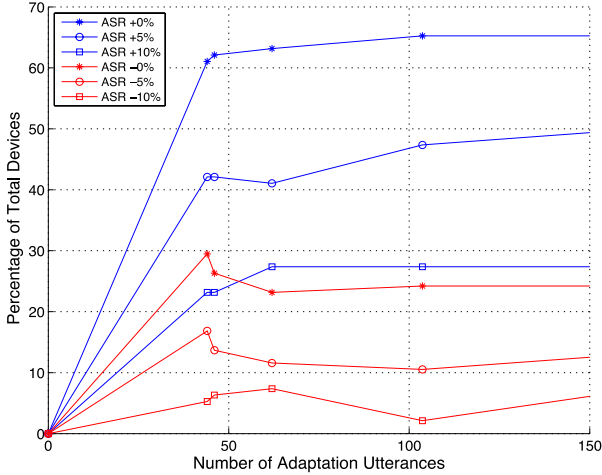


Figure 4: Percentage of devices experiencing increase/decrease in word error rate as a function of the number of adaption utterances in unsupervised learning.

4.2 Language Model Adaptation

Large performance improvements are possible when domain specific language information is used. In mobile communication, the challenge when using DS information is capturing sufficient data to enable adaptive learning of language models. In our study, the use of HLMs provided an ideal framework for unsupervised learning even when limited data is available. The structure of the HLM enables us to adapt selected constituents depending on the availability of data.

In this section, we adapted the HLM by exploiting DS information from the contact list stored on the device itself. Here we compare the speech recognition performance on utterances spoken into the home screen widget. There are two comparisons using the same acoustic model. In the first we compare the performance on all the utterances from the home screen using two different language models. One is trained on all the available data with a large list of DI contact information, and the second is employing DS contact list information. The test set consists of 7444 utterances and 42240 words. The performance is shown in Table 1.

	Sub	Del	Ins	Err
DI	13.7	2.5	3.3	19.5
DS	8.7	2.3	2.3	13.3

Table 1: Performance comparison when the knowledge of the device contact list is available to the HLM.

It is possible to single out utterances that are more likely to benefit from the knowledge of the device contact list. Here it is done by selecting all the utterances that contain any of the following words: *send, text, email, call, dial, message and phone*. Those words are present in 4958 out of the original 7444 utterances and they contain 32130 reference words.

Table 2 shows that the recognition performance improvement is even more significant for this subset of utterances.

	Substitution	Deletion	Insertion	Error
DI	14.1	2.3	3.2	19.6
DS	7.6	2.1	1.9	11.6

Table 2: Performance comparison when the knowledge of the device contact list is available to the HLM in the relevant subset of utterances.

Tables 1 and 2 show significant reductions in word error rate when the HLM is adapted with DS contact list information. For the relevant subset of utterances, we obtained an absolute reduction of 8% in word error rate.

5. Summary

This paper reviewed our research towards developing advanced mobile virtual assistants. In particular, we presented our acoustic and language-modeling algorithms for supervised and unsupervised adaptation of MVAs. Our results showed that we are able to achieve improvement in acoustic model adaptation with only 20 utterances of device specific data, and that with 40-50 utterances, unsupervised learning converges to that of supervised learning. In a related study, we showed that 65% of devices experienced a gain in accuracy with unsupervised adaptation. For hierarchical language modeling, we showed that device specific information results in 8% absolute reduction in word error rate.

References

- [1] Bocchieri, E. and Caseiro, D., "Use of Geographical Meta-Data in ASR Language and Acoustic Models." *Proc. Int. Conf. Acoustics Speech, and Signal Processing*, Dallas, 2010.
- [2] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar. "The AT&T watson speech recognizer." *Proc. Int. Conf. Acoustics Speech, and Signal Processing*, Philadelphia, 2005.
- [3] Brown, M. K. and Glinski, S. C., "Context-free large vocabulary connected speech recognition." *Proc. Int. Conf. Acoustics Speech, and Signal Processing*, Adelaide, 1994.
- [4] Galescu, L. and Allen, J., "Hierarchical Statistical Language models Experiments on In-domain Adaptation." *Proc Int. Conf. Speech and Language Processing*, Beijing, 2000.
- [5] Bacchiani, M., Riley, M., and Roark, B., "MAP adaptation of stochastic grammars", *Computer Speech and Language*, 2006.
- [6] Young, S.J. Odell, J.J., and Woodland, P.C., "Tree-Based State Tying for High Accuracy Acoustic Modeling," *Proc. ARPA HLT Workshop*, pp. 307-312, Morgan Kaufmann, 1994.
- [7] Gauvain, J.L., and Lee, C.H., "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains." *IEEE Trans. Speech and Audio Proc.*, pp 291-298, vol 2, 1994.
- [8] Bocchieri, E., Riley, M., and Saraclar, M., "Methods for Task Adaptation with Limited Transcribed in-Domain Data." *Proc Int. Conf. Speech and Language Processing*, South Korea, 2004.
- [9] Gales, M.J.F., "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, 12:75-98, 1998.
- [10] Shinoda, K., and Lee, C.H., "Structural MAP Speaker Adaptation Using Hierarchical Priors." *Proc. IEEE-ASRU Workshop*, 1997.
- [11] Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., "A Compact Model for Speaker adaptive Training", *Proc. Proc Int. Conf. Speech and Language Processing*, 1996.
- [12] Tur, G., Wright, J., Gorin, A., Riccardi, G., Hakkani-Tu, D. "Improving spoken language understanding using word confusion networks." *Proc. of the International Conference on Spoken Language Processing*. Denver, 2002.