# SPEECH RECOGNITION MODELING ADVANCES FOR MOBILE VOICE SEARCH

*Enrico Bocchieri, Diamantino Caseiro, and Dimitrios Dimitriadis*

*AT&T* Research, 180 Park Ave, Florham Park, New Jersey 07932

## ABSTRACT

This paper reports on the development and advances in automatic speech recognition for the *AT&T* Speak4it® voice-search application. With Speak4it as real-life example, we show the effectiveness of acoustic model (*AM*) and language model (*LM*) estimation (adaptation and training) on relatively small amounts of application field-data. We then introduce algorithmic improvements concerning the use of sentence length in LM, of non-contextual features in AM decision-trees, and of the Teager energy in the acoustic front-end. The combination of these algorithms, integrated into the *AT&T* Watson recognizer, yields substantial accuracy improvements. LM and AM estimation on field-data samples increases the word accuracy from 66.4% to 77.1%, a relative word error reduction of 32%. The algorithmic improvements increase the accuracy to 79.7%, an additional 11.3% relative error reduction.

***Index Terms***— speech recognition, HMM, decision tree clustering.

## 1. INTRODUCTION

Several applications of voice-search for business listings have recently been developed for cell-phones [1, 2]. Speak4it® is a client-server application developed at *AT&T* for the iPhone®, and available from the Apple® Store. The user freely says in one turn the description, and optionally location, of the desired business, to access the business information presented on the phone display as a map with hyper-text links. This paper presents the development and algorithmic advances of the large vocabulary speech recognizer (*AT&T* Watson recognizer [3]) for the Speak4it task.

The estimation of acoustic (*AM*) and language (*LM*) models on application field-data has proved very important for achieving high recognition accuracy: field-data are useful even in relatively small amounts compared to the data for training task-independent models.

Section 2 describes the *pilot* system for the initial deployment, and Section 3 the system upgrades and recognition accuracy improvements by the AM and LM estimation on temporal snapshots of field-data. Section 4 presents improvements in modeling, concerning a LM conditioned on sentence length, the use of non-contextual features in AM decision trees, and improvements in robustness of the acoustic front-end. The recognition accuracy is significantly improved by the combination of these methods, as summarized in Sections 4.4 and 5.

## 2. PILOT ASR SYSTEM

Application client prototypes were installed on iPhone units, and used by *AT&T* employees to record speech during daily activities. Due to the lack of geographical (New Jersey and California) and demographic coverage, we decided against adapting models on these data. They were used mostly as a development set to evaluate the tuning of the pilot recognizer.

### 2.1. Language Model

A Katz backoff 3-gram LM was built using 10 million anonymous business search queries from *yellowpages.com*. These are typed, *not spoken*, queries consisting of two fields: search term (e.g. business name), and location. To approximate the text of spoken queries for LM estimation, these fields were combined in one sentence using the most frequent carrier phrases in the development data, weighted by relative frequency. Misspellings and inconsistent business name tokenizations in the user-typed web queries were corrected by $62k$ substitution rules. After text normalization, the vocabulary contains $240k$ words.

### 2.2. Acoustic Model

The "general" telephone acoustic model of the *AT&T* recognizer is a discriminatively trained tri-phone hidden Markov model (*HMM*), with states consisting of Gaussian mixtures, estimated on $2k$ hours of telephone audio data. The acoustic frames have 60 dimensions, defined by discriminative projections of MFCC vectors [3]. To better match the AM to the iPhone recordings, we adapted the AM on 120 hours of wireless speech data from a directory assistance (*DA*) application, by the maximum a posteriori (*MAP*) method. The MAP prior was set to optimize the iPhone pilot data recognition accuracy. This prior value weights the DA data for half of the total count.

## 3. DEPLOYMENT, AND SYSTEM UPDATES

The pilot recognizer of the previous Section was deployed in Dec. '08, and the word accuracy in the first week was $\approx 65\%$, i.e. 15% lower than in the pilot study. It was immediately

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

evident, and expected, that real users talk in noisier environments, following different patterns than the "pilot" users: for example the city-state location (relatively easy to recognize correctly) is not specified in 80% of the field-data, and many sentences are not business queries (see Section 4.1). To improve the recognizer performance, a fraction of the recordings from the field was transcribed (high quality transcriptions of $\approx$ 1,000 queries/day, with average 2.5 words/query) for model adaptation and training. In this paper, we refer to four different Speak4it recognition systems as follows:

***r1*** : pilot LM and AM, no field-data available.
***r2*** : LM and AM upgrades based on $30k$ spoken queries.
***r3*** : ” ” $158k$ spoken queries.
***r4*** : ” ” $337k$ spoken queries.

### 3.1. LM Adaptation

In general, we have used interpolation to combine several models estimated on the field-data and on the much larger and more general web-data. Interpolation provides us with free parameters, namely the interpolation weights, that we estimate by minimizing the LM perplexity of held-out field data.

First, we re-estimated the frequency of carrier phrases on the field data and we rebuilt the web-queries language model. Then, this model was linearly interpolated with 2 and 3-gram models estimated from the available field transcriptions, and with several models estimated on subsets of field-data and web-data, respectively.

The strategy for selecting data subsets changed as more queries were transcribed. In ***r2***, the web corpus was partitioned in 200 subsets using k-means clustering of td-idf weighted term vectors, and the cosine distance. In ***r3***, we divided the field data in weekly data sets. Since we got diminishing improvements by partitioning the data as the field data increased, we decided in ***r4*** to use only the complete web and field data models.

### 3.2. AM Adaptation and Re-Training

The adaptation of the ***r1*** acoustic model on the transcribed field-data sets was performed by:

- the application of *MLLR* to the Gaussian means, then
- two iterations of supervised segmentation and MAP adaptation of Gaussian means, variances and mixture weights.

This process generates several adapted models for different MAP prior values, respectively, and the best model is selected by testing on held-out data. The AM adaptation is automatic, given the speech data and their transcriptions. The *AT&T* Natural Voices$^{®}$ text-to-speech system provides the dictionary entries of the words in the transcriptions. In general, AM's trained on task-specific data yield higher recognition accuracy than AM's adapted on the same data, when sufficient data are available: for the Speak4it task, this corresponds to $110k$ (or more) queries.

Thus, the ***r2*** AM is derived by adaptation of the ***r1*** AM (on $30k$ queries). By contrast, the AM's ***r3*** and ***r4*** are fully trained AM's on $158k$ and $337k$ queries, respectively. Their sizes are:

***r1***, ***r2*** AM's: $44k$ triphones, $14k$ states, $278k$ Gaussians.
***r3*** AM: $14k$ triphones, $4.8k$ states, $96k$ Gaussians.
***r4*** AM: $19k$ triphones, $7.6k$ states, $153k$ Gaussians.

The ***r1*** and ***r2*** AM's have the size of the generic AM of Section 2.2, from which they derive by adaptation steps. The ***r3*** and ***r4*** AM's are smaller because they are trained on the smaller field-data sets, and their sizes were optimized by experiment.

### 3.3. Accuracy Improvements

The accuracies (6,461 queries test-set) of systems ***r1*** through ***r4*** are plotted in Fig. 1 versus the *real-time factor* (CPU time divided by the input speech duration). The accuracy curves are obtained by varying the pruning beam-width of the Viterbi decoder. All other computational speed-up techniques, such as Gaussian selection and state histogram pruning, are turned off, since the emphasis of the paper is on ASR accuracy. The asymptotic word accuracies of systems ***r1*** through ***r4*** are also presented in the first four lines of Table 1, respectively.

The improvements of ***r2***, ***r3*** and ***r4*** w.r.t. ***r1*** reflect both the LM and AM upgrades. The respective asymptotic accuracies of ***r1*** and ***r4*** are 66.4% and 77.1%, for a 10.7% absolute word error rate reduction. There is often interest in comparing the LM and AM upgrades: in our opinion, their relative effectiveness is determined by factors such as the available data, and suitability of the base-line models.

When using the largest field-data set, as in ***r4***, upgrading only the AM works better than upgrading only the LM, with accuracies of 72.9% and 70.6%, respectively.

However, the opposite held with fewer field-data, as in ***r2***: the adaptation of both the LM and AM increases the asymptotic word accuracy to 71.6%, while the adaptation of just the LM or AM improved the asymptotic word accuracy to 69.9% and 67.9% respectively. This was so because the ***r1*** LM was highly mis-matched, and its adaptation on even the small amount of field data (***r2***) proved sufficient in learning the statistics of few carrier phrases (Sections 2.1, 3.1). However, additional LM adaptation in ***r3*** and ***r4*** provides smaller improvements of accuracy, as very large data are generally needed for the estimation/adaptation of statistical LM's.

## 4. ALGORITHMIC IMPROVEMENTS

In this Section, we present algorithmic improvements of the LM, AM and acoustic front-end. The proposed LM and AM are trained on the ***r4*** field data set, thus the ***r4*** recognizer serves as baseline.

### 4.1. Sentence Length Conditioned LM

Analysing the field-data transcriptions we observed that about 10% of the input queries are out-of-domain. The user often tries to use speech for unsupported functionalities (e.g. map navigation or voice dialing), or simply plays with the application uttering arbitrary speech. Since these out-of-domain sentences are often longer than in-domain utterances, and clearly correspond to a different user intent, we decided to condition the LM estimation on sentence length. We built a 3-gram language model for each of six subsets of the transcribed training data, containing, respectively sentences of length $|s| \leq 3, |s| = 4, |s| = 5, |s| = 6, |s| = 7, |s| > 7$. These language models were linearly interpolated with a 3-gram language model trained on all transcription and with the web query model, by minimizing the perplexity on a development set.

### 4.2. Decision Trees for Context Modeling

It is common practice in HMM-based ASR systems to train *context-dependent* phonetic HMM's, with state observation p.d.f.'s conditioned on the phonetic context. To solve data sparsity problems, the phonetic contexts are clustered by applying decision trees to the training data of the HMM monophone states [4], a process also known as state tying.

Our decision-tree algorithm is based on the greedy maximization of the likelihood of Gaussian models with diagonal covariance (a known method not reviewed here), exploiting sufficient statistics for computational speed in training. In our systems ***r1*** through ***r4***, tree-node splitting is based on binary questions on the phonetic context, characterized by the features of place and manner of articulation. The phonetic context (±1 for triphones and ±2 for pentaphones) spans across word boundaries. To improve the recognition accuracy, we have recently introduced:

i. Decorrelation of training vectors for every tree, to approximate full-covariance Gaussian models.

ii. Binary questions about phonetic context expressed as membership to phoneme sets (now 87, including sets of only one phoneme).

In experiments on Speak4it, pentaphone AM's have shown small accuracy improvements over triphones (see also [5]), with longer decoding times. Thus, we prefer to improve triphone models by means of:

iii. features other than the phonetic context, encoded by adding tags to monophonic labels [6]. Differently tagged monophones share the same decision tree where nodes may be split according to binary properties of both the tags and the phonetic context.

The motivation is to condition the phonetic HMM's on meta-information such as "non-native speaker" , without *forcing* the HMM state estimation on sparse subsets of data.

We call *lexical features* those, such as *word* and *syllable* [7], handled by modifications of dictionary entries during AM training and recognition. For example, we tag phoneme $/p/$ with "**F**" when it is word *final*. The $/p/$ and $/p_F/$ labels are contained in the sets that define the tree-building contextual questions (ii above) The $/p/$ and $/p_F/$ labels share the same decision tree (iii above), and tree nodes may be split by the phonetic tag and by phonetic context.

We label entire utterances with *sentence* ("**S**") features that typically refer to *classes* of speakers (e.g. gender, foreign-accent), or of acoustic environments [8]. Tree building can proceed as with the *lexical* features, simply by applying the tag to the phonemes of whole utterances. However decoding is different, because the input sentences are not tagged. Tagging could be performed by a sentence classifier (two-pass approach), introducing, however, latency in our real-time application. Therefore we decode the input sentence for the different tag values *simultaneously* by the same decoder: thanks to the beam-search pruning the time overhead becomes small.

We think it is very useful to define features automatically from data, instead of requiring hand-labelling of the training sentences, as needed for gender and dialect [8]. Our idea is to "seed" binary sentence classifiers with the acoustic frames from small data sets, and to iterate the classifier estimation on unlabelled training sentences: then, the classifier outputs become binary features of the training sentences (others have studied the automatic generation of phonetic contextual questions [9], not of features). As a first approach, our seed data consists of $10k$ sentences labelled by speaker gender to bootstrap a Gaussian-mixture gender classifier (either "man" or "woman/child"). This is ultimately applied to the ***r4*** $337k$ training sentences, to provide the "**S**" tags for tree growing.

### 4.3. TECC Front-End

The acoustic front-end of the *AT&T* Watson recognizer [3] is based on the mel-frequency cepstral analysis of speech. A vector of 21 cepstral coefficients (*MFCC*) is computed for each frame, and mean-normalization is applied in real-time. To capture the speech signal dynamics, 11 consecutive vectors are concatenated into a super-vector, and then projected onto a 60 dimensional feature space. The projection combines a discriminative feature extraction known as heteroschedastic discriminant analysis (*HDA*), and a decorrelating linear transformation, i.e. MLLT.

It has been shown that the Teager-Kaiser energy estimate [10, 11] is more robust than mean-squares, since it presents less prominent transient phenomena and smaller deviation errors in certain noisy conditions. Therefore we have substituted the Teager energy cepstral coefficients (*TECC*) in place of the MFCC's in the recognizer front-end. The HDA and MLLT projections have also been estimated for the TECC's, see [10]. The TECC based front-end is more accurate than the MFCC's, and the improvements from the better LM, AM and front-end add up nicely, as discussed next.

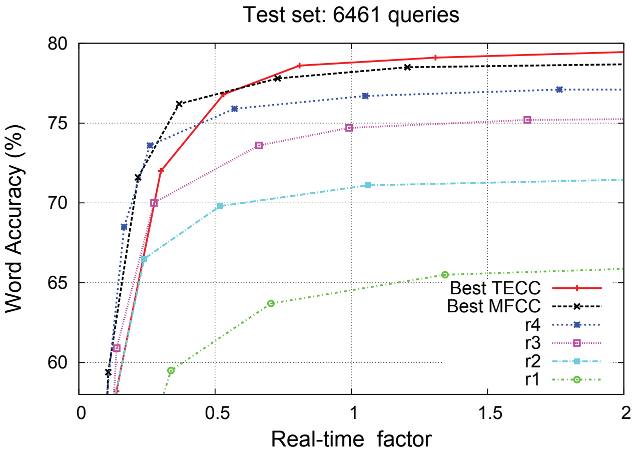

**Figure 1**. ASR performance of different Speak4it systems.

### 4.4. Results

The accuracy of 77.1% of the ***r4*** system ($4^{th}$ line of Table 1) is the baseline for the algorithm improvements of this Section. The lower portion of Table 1 (below the $4^{th}$ line) shows the accuracies for the various improvements, with the LM and AM being estimated on the same field-data as ***r4***. To assess the relative importance of the improvements, the lower part of Table 1 contains lines that differ only by one system component (LM, AM or front-end). In the LM column, "*SL*" refers to the "sentence length" LM (Section 4.1), with 0.4% accuracy increase. In the AM column $i$,$ii$, and $iii$ refer to the respective decision-tree upgrades (Section 3.2), with tags "**F**" and "**S**" for the non-contextual features. All the AM's are triphonic except when denoted by "$5ph$" (pentaphones). The pentaphones improve the word accuracy by only 0.2%, which prompted us to use non-contextual features ($iii$).

The combination of the AM and LM improvements increase the accuracy by 1.8%, from 77.1% to 78.9%, without changing the MFCC frontend. Switching from the MFCC to the TECC frontend improves the accuracy further to 79.7%. The overall relative word error rate reduction is 11.3%. In Fig. 1, the two top curves correspond to the best TECC and MFCC systems, demonstrating that the improvements hold at real-time.

## 5. DISCUSSION

Using the Speak4it application as a real-life example, we have shown in Section 3 the importance of AM and LM adaptation and training on even relatively small amounts of field-data, improving the word recognition accuracy from 66.4% to 77.1%, for a relative word error rate reduction of 32%. Section 4 further improves the accuracy to 79.7%, with an additional relative error reduction of 11.3%, by means of combined algorithmic improvements, concerning sentence-length conditioned LM, the use of non-contextual features in AM decision-trees, and acoustic front-end robustness.

| LM | AM | Front-End | Word Acc. (%) |
|---|---|---|---|
| ***r1*** | ***r1*** | *MFCC* | 66.4 |
| ***r2*** | ***r2*** | *MFCC* | 71.6 |
| ***r3*** | ***r3*** | *MFCC* | 75.4 |
| ***r4*** | ***r4*** | *MFCC* | 77.1 |
| ***r4*** | $i, ii$ | *MFCC* | 77.6 |
| ***r4*** | $i, ii, 5ph$ | *MFCC* | 77.8 |
| ***r4*** | $i, ii, iii_F$ | *MFCC* | 77.9 |
| *SL* | $i, ii, iii_F$ | *MFCC* | 78.3 |
| *SL* | $i, ii, iii_S$ | *MFCC* | 78.9 |
| *SL* | $i, ii, iii_F$ | *TECC* | 79.0 |
| *SL* | $i, ii, iii_S$ | *TECC* | 79.7 |

**Table 1**. Word accuracies (asymptotic at large beam-widths) for various system components

Interestingly, the introduction of non-contextual features does not change significantly the AM size, but it allows for splitting the tree-nodes by other ways than phonetic context. In Table 1 the "**S**" feature yields a larger accuracy gain than "**F**": this correlates with the structure of the decision-trees, where the questions about "**S**" and "**F**" are applied for growing 90.5% and 11% of the tree-leaves (HMM states), respectively. We are continuining our investigation on non-contextual features for AM, and on robust acoustic front-ends.

## 6. REFERENCES

[1] C. van Heerden et.al., "Language modeling for What-with-Where on GOOG-411," *Proc. of INTERSPEECH*, 2009.

[2] A.Acero et.al., "Live search for mobile:Web services by voice on the cellphone," *Proc. of ICASSP*, pp. 5256–5259, 2008.

[3] V.Goffin et al., "The AT&T Watson Speech Recognizer," *Proc. of ICASSP*, 2005.

[4] J.J.Odell S.J.Young and P.C.Woodland., "Tree-based state tying for high accuracy acoustic modelling," *Proc. of H.L.T.*, pp. 307–312, 1994.

[5] D.Rybach and M.Riley, "Direct construction of compact context-dependency transducers from data," *Proc. of INTERSPEECH*, pp. 218–221, 2010.

[6] D.B. Paul, "Extensions to phone-state decision-tree clustering: single tree and tagged clustering," *Proc. of ICASSP*, pp. 1487–1490 vol.2, 1997.

[7] H.Liao C.Alberti M.Bacchiani and O.Siohan, "Decision tree state clustering with word and syllable features," *Proc. of INTERSPEECH*, 2010.

[8] C.Fügen and I.Rogina, "Integrating dynamic speech modalities into context decision trees," *Proc. of ICASSP*, 2000.

[9] K.Beulen and H.Ney, "Automatic question generation for decision tree based state tying," *Proc. of ICASSP*, 1998.

[10] D. Dimitriadis, E. Bocchieri and D. Caseiro, "An alternative front-end for the AT&T Watson LV-CSR system," *in Proc. of ICASSP-11*, 2011.

[11] D. Dimitriadis, P. Maragos and A. Potamianos, "On the effects of filterbank design and energy computation on robust speech recognition," *accepted in Trans. on Audio, Speech and Language Processing*, 2010.