# Improve Speech Recognition in Mobile Environment by Dynamical Noise Model Adaptation

*Changxue Ma, Yanjun Wei*

Human Interface Lab.
Motorola Labs
1301 E. Algonquin Rd, Schaumburg, IL 60196
{Changxue.Ma, Yanjun.Wei}@Motorola.com

## Abstract

In this paper, we identify that the major source of error in mobile device based speech recognition is the long silence segment in the beginning and the end of the utterance, which is most noise sensitive. While these segments are most dynamic, they are often poorly modeled by HMM. To improve the silence modeling could improve the overall performance of the speech recognition systems. We propose in this paper to dynamically adapt the silence model on the utterance level. Using a multi-lingual database collected in car under a number of driving conditions, we show a significant improvement in speech recognition accuracy.

## 1. Introduction

The performance of speech recognition systems degrades significantly when mismatched conditions occur where the training samples and testing samples are taken from different acoustical environments. Sophisticated techniques including frond-end signal processing and HMM models adaptations have been proposed to compensate for the mismatches [2][3][4]. For the database collected in noisy environments, the matched condition training and testing are often preferred compared with situations where the sophisticated noise reduction techniques are used to reduce mismatches. But the matched conditions are hard to obtain due to the difficulty of data collection and diversity of changes in speakers and acoustical environments. The speech recognitions for speech in mobile environment are such a challenging task where the noise is often introduced and acoustical environment is dynamic, especially in car-driving conditions.

When acoustical environment changes, the silence model or non-phoneme model needs most attention. Firstly, duration modeling in silence can be seriously deficient in the presence of long non-speech signal. Unlike phonemes, whose durations in the utterances are, in some degree, constrained by the physics of movement of the vocal apparatus, there are no physical constraints on the duration of a silence. Further more it is well known that duration modeling in HMM is deficient. Secondly, the silence part of the utterance is most noise sensitive. Therefore, in this paper, we will focus on the silence HMM model adaptation to improve the recognition performance in noisy conditions.

HMM model adaptation techniques like Maximum a Posteriori (MAP) estimation and Maximum likelihood linear regression (MLLR), mostly used for speaker adaptation, are less effective when the data is scarce, in particular, in on-line situation [1][2]. Even HMM adaptation with MLLR needs a number of adaptation utterance [1]. The scarceness of data can leads to a wrong or poor estimate of the linear transformation. This shared transformation could then damage the underlying structure of the whole acoustic space. It can be worse for noisy speech under unsupervised situations. In the same token, we can also view the global Cepstral Mean Subtraction (CMS) technique as an on-line adaptation technique, where the all means of the Gaussian mixtures is adjusted based on the cepstral means of the recognizing utterance.

In these adaptation schemes, all Gaussian mixtures in the HMM models can be modified. We know that HMM models with multi-mixture are typical setup to handle the diversity of the speech data. In general, a greater diversity in speakers and acoustical environments requires more mixtures in HMM models. So speaker dependent speech recognition generally outperforms the speaker independent speech recognition with the same HMM setup. In other words, to achieve the same performance for both situations, HMM models with less number of Gaussian distributions can be used. For each particular utterance, we basically perform speaker dependent recognition if we can adapt the model with the on-line data. This implies that for each utterance only limited number of mixture components affects the outcome of score ranks. This leads us to propose to update selected mixtures on-line.

In this paper, we will show that one of the major sources of errors in mobile device based speech recognition is the long and noise corrupted silence segment in the beginning and the end of the utterance. While these segments are most dynamic and are poorly modeled by HMM, we propose to instantly adapt the silence model on the utterance level. Using a multi-lingual database collected in car under a number of driving conditions, we show a significant improvement in speech recognition accuracy

## 2. Multilingual Digit Database Collected in Cars

### 2.1. Data collection

As a part of the AURORA project, Speech-dat Car (SDC) digits database has been established for Italian, Danish, German, Spanish and Finnish [5]. All sessions have been recorded with native speakers using both a close-talking (CT) microphone and a hands-free (HF) microphone in cars under five different driving conditions:

| **Driving condition** | **Condition** |
|---|---|
| 0 km/h, engine on | Quiet |
| 40-60 km/h | Low noise |
| 40-60 km/h, window open | Low noise |
| 100-120 km/h | High noise |
| 100-120 km/h, music on | High noise |

The databases are organized into three different matching conditions:

- Weakly mismatch condition (WM) where training uses 70% of data from both HF and CT microphones and testing uses the rest 30% data from HF and CT microphones in all driving conditions.
- Medium-mismatch condition (MM) where training uses 70% of *quiet* and *low noise data* from HF microphone and testing uses 30% of *high noise data* from HF microphone only.
- High-mismatch condition (HM) where training takes 70% of all conditions from CT microphone and testing uses 30% of *low noise* and *high noise* data from HF microphone.

For each test, speakers are split into separate training and test sets so that 70% of the speakers are in the training set and the rest in the test set.

### 2.2. Feature generation

A cepstral analysis scheme recommended in the Aurora WI007 front-end is performed [5]. Speech signal offset is first compensated with a notch filtering operation. We use a frame length of 25ms speech with a Hamming window and the frame advance of 10 ms. The pre-emphasis factor of 0.97 is applied. FFT based Mel frequency filter-bank analysis produces 23 frequency bands in the range from 64 Hz up to half of the sampling frequency. Finally all 13 Mel frequency cepstral coefficients (MFCCs) are generated. The reference baseline results and our experiments are based on the same features.

### 2.3. Acoustic modeling.

In the Aurora project specification, the HMM models are whole word digit model and their topologies are the same for the five languages. The silence HMM model has three emitting states and each state contains a mixture of six Gaussian distributions. The pause HMM model has a single state that is tied to the middle state of the silence model. Each digit HMM model has sixteen states and each state contains a mixture of three Gaussian distributions. The training procedure is the same for all five languages. For testing, the recognition grammar and language model penalty is also the same for all five digit-recognition tasks.

## 3. End-silence segment removal with forced alignment

It is observed that the silence removal can significantly improve the speech recognition performance. That is why good end-pointing algorithm plays an important role in real life speech recognitions. We intend to show here significant accuracy improvement can be achieved by adequately delete silences.

For the purpose of producing the accurate alignment, five sets of HMM models are trained with data recorded from the closed-talking microphone so that high recognition accuracy maintains. The alignment then is performed on the test data recorded also from close-talking microphones. The alignment information is used for removing the silence segments in the beginning and end of the corresponding utterances recorded in the hands-free and close-talking modes. For silence segments in the beginning of the utterance, we removed the silence frames up to the *last five frames*. For silence segments in the end of the utterance, we keep only *five frames in the beginning of a silence segment and remove the rest.* As a result, we then have prepared a version of testing database without ending-silence segments. We follow the same procedure for training and testing as proposed in [5]. We indeed achieved the performance improvement as shown in Table 1 and 2. The weighed average error rate reduction is across five languages is 17.31%.

*Table 1. Performance of standard AURORA project setup with ending-silence deletion.*

| Training Mode | Seen Databases | | | Unseen Databases | | Average |
|---|---|---|---|---|---|---|
| | Italian | Finnish | Spanish | German | Danish | |
| Weakly Mismatched | 94.90 | 93.23 | 88.28 | 91.02 | 84.62 | 90.41 |
| Medium Mismatch | 84.38 | 82.15 | 76.93 | 78.70 | 63.41 | 77.11 |
| High Mismatch | 55.07 | 48.59 | 51.46 | 74.56 | 48.21 | 55.58 |
| 0.4W+0.35M+0.25H | 81.26 | 78.19 | 75.10 | 82.59 | 68.09 | **77.05** |

*Table 2: Performance improvement over the baseline results due to silence segment reduction.*

| Training Mode | Seen Databases | | | Unseen Databases | | Average |
|---|---|---|---|---|---|---|
| | Italian | Finnish | Spanish | German | Danish | |
| Weakly Mismatched | 19.81 | 28.51 | 10.87 | 4.67 | 22.32 | 17.24 |
| Medium Mismatch | 13.13 | 35.09 | 12.15 | -1.72 | 25.07 | 16.74 |
| High Mismatch | 25.32 | 26.19 | 15.98 | 1.09 | 22.62 | 18.24 |
| 0.4W+0.35M+0.25H | 18.85 | 30.23 | 12.60 | 1.54 | 23.36 | **17.31** |

## 4. Noise model adaptation

As we have demonstrated in the section above, noise segments both in the beginning and the end can be a major source of recognition degradation for telephone-like speech. Intra-word silences can have similar adverse effects on recognition results. As we have states in the above that for each utterance only limited number of mixture components affects the outcome of score ranks. Minor degradation incurs when mixture calculation takes the value of the maximum of the mixture components. This leads us to propose to update selected mixtures on-line. If we can select in some way in advance those mixture components that participate the score ranking operation, we can only update those components while maintain other components unchanged. We propose here the *reduced model adaptation*. The reduced model contains less mixtures and adaptation will be more robust. This adaptation will take place in the decoding process.

Procedure for *reduced model adaptation*

1) Calculate the average of the first n frames as:

$$M_s = {}^{1}\!/_{n} \sum_{i=0}^{n-1} O_i$$

2) For each state of the silence model, identify the target mixture component as:

$$\mu_{il} = Max_{j \in (gaussians)} N_{ij}(M_s, \mu_{ij}, \Sigma_{ij})$$

3) Update selected mean as:

$$\mu_{il} <= (1-\alpha) * u_{il} + \alpha * M_s$$

4) Restore the original mean vector $\mu_{il}$ for the next utterance.

As we can see, this method produce very little overhead in terms of both CPU and memory consumption.

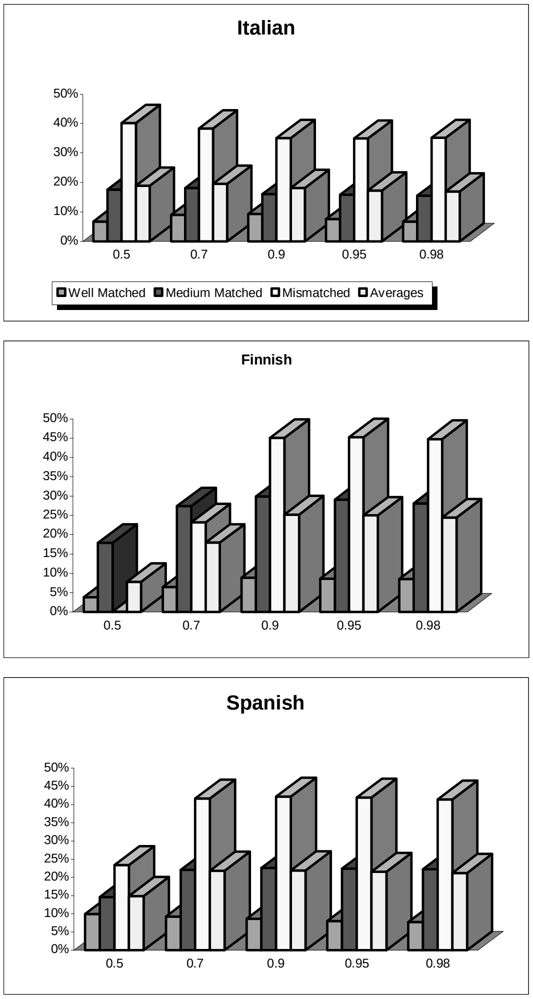
## 5. Experiment Results

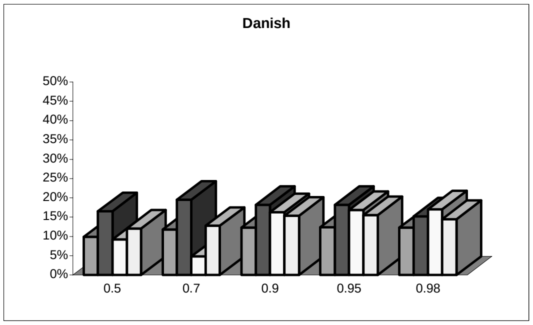
### 5.1. Performance improvement as a function of updating weight.

In this experiment, we test the influence of the updating weight factor α on the speech recognition performance. The number of frames used to obtain mean is fixed as 10. Figure 1 and Figure 2 show the error rate reduction for five languages as a function of α.

We see the performance is stabilized around a α value of 0.9. The German and Danish languages listed as unseen database, are shown less improvement. One of the possible reasons could be that HMM structure and recognition setup was experimentally chosen based on the other three languages. Much more deletion errors than insertion errors are seen in these two databases.

*Figure 1: Error rate reduction as a function of updating factor for five languages.*





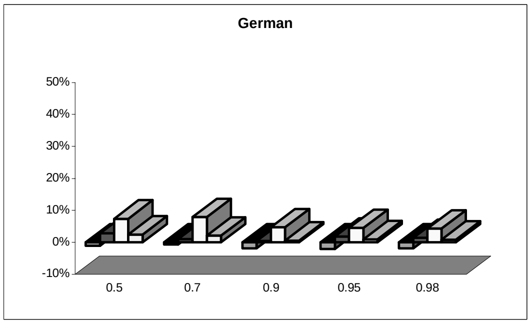
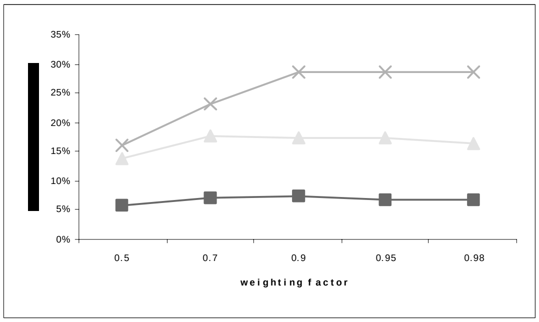


*Figure 2: Average error rate reduction among five languages in three matching conditions. HM is on the top, WM on the bottom, and the middle line is for the MM condition.*



### 5.2. Update all the means in the silence model.

In this test, we update all the silence means. We found that when all mixture is update with the new mean, the improvement is significantly lower, for example, in the case of alpha = 0.7, error rate reduction decreased from averaged 14.82% to a 11.14 %.

### 5.3. Silence model adaptation after Cepstral Mean subtraction

We also test the situation when cepstral mean subtraction is used. Because the initial conditions for calculating the dynamic cepstral coefficients are always problematic, in this experiment, we also show the difference in the performance between updating all 39 cepstral coefficients and just updating the 13 static cepstral coefficients. However, the difference is just marginal. We can see from Table 3 and 4 the strong contributions from CMS techniques and *reduce model adaptation.*

*Table 3: Error rate reduction with silence model adaptation and CMS. (39 cepstral coefficients)*

| Training Mode | Seen Databases | | | Unseen Databases | | Average |
|---|---|---|---|---|---|---|
| | Italian | Finnish | Spanish | German | Danish | |
| Weakly Mismatch | 14.78 | 41.71 | 19.09 | 9.84 | 23.28 | 21.74 |
| Medium Mismatch | 28.92 | 52.00 | 40.56 | 11.22 | 13.07 | 29.15 |
| High Mismatch | 47.17 | 52.82 | 46.11 | 32.74 | 19.68 | 39.70 |
| 0.4W+0.35M+0.25H | 27.83 | 48.09 | 33.36 | 16.05 | 18.81 | **28.83** |

*Table 4: Error rate reduction with silence model adaptation and CMS. (13 cepstral coefficients)*

| Training Mode | Seen Databases | | | Unseen Databases | | Average |
|---|---|---|---|---|---|---|
| | Italian | Finnish | Spanish | German | Danish | |
| Weakly Mismatch | 16.82 | 42.34 | 19.39 | 10.27 | 24.75 | 22.72 |
| Medium Mismatch | 28.92 | 51.75 | 40.25 | 10.51 | 16.81 | 29.65 |
| High Mismatch | 45.86 | 52.06 | 46.11 | 30.40 | 20.17 | 38.92 |
| 0.4W+0.35M+0.25H | 28.32 | 48.06 | 33.37 | 15.39 | 20.83 | **29.19** |

## 6. Conclusions

We propose to adapt the selected mixtures in the silence model to improve the robustness of speech recognition in noisy environment. This approach shows the biggest error rate reduction in highly mismatch testing conditions. Compared with other noise reduction techniques, the reduced silence model adaptation is more cost-effective.

## 7. References

[1] Lee, C.-H., Lin, C.-H. and Jung, B.H., "A study on speaker adaptation of continuous density HMM parameters," ICASSP, 1990, pp.145-148.

[2] Deng, L., Acero, A. Plumpe, M., and Hung, X.D., "Large-vocabulary speech Recognition under Adverse Acoustic Environments", Proc. *ICSLP, Vol. III, 2000.*

[3] Boll, S., "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoustics, speech and Signal Proc., Vol. 27, 1979, pp. 114-120.

**[4]** Gong, Y., and Godfrey, J,J., "Transforming HMMS for Speaker-Independent Hands-Free Speech Recognition in The Car", ICASSP, 1999, pp. 297-300.

**[5]** H G Hirsch & D Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"; Paris, France, September 18-20, 2000