Adapting Acoustic Models for Speech Recognition

Kalin Stefanov and Raveesh Meena
KTH Royal Institute of Technology
kalins@kth.se,raveesh@kth.se

June, 2014

Abstract

Variability in the environment (noise, room acoustics, distance from microphone, type of microphone) is a challenge for automatic speech recognition. In this paper, we propose and evaluate a scheme for adapting the acoustic model of an ASR system to address this challenge. The key idea is to transform the data (speech recordings) used for training the acoustic models, as if it was recorded in the target environment. This involves, first to model the acoustics of the target environment and then, transform the original speech recordings to reflect the characteristics of that environment, and finally, train acoustic models on the transformed data. An evaluation shows that it is in fact possible to obtain gains in recognition performance following the proposed scheme.

1 Introduction

The task of a speech recognition system is to provide a sequence of words that best match given speech sounds. Figure 1 illustrates the typical components of modern ASR systems.

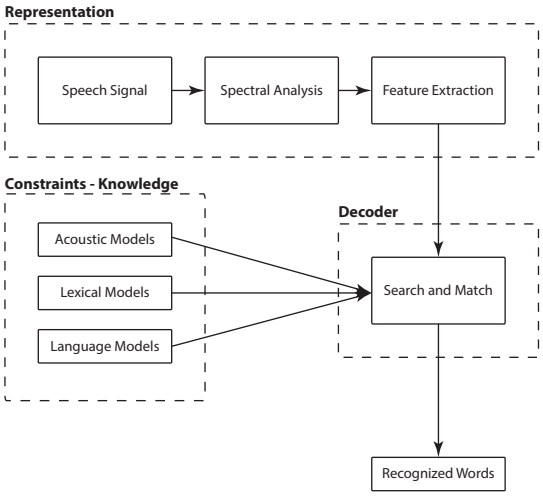

Figure 1: Modern ASR system

From probabilistic perspective speech recognition systems attempt to solve,

$$P(words|sounds) = \frac{P(sounds|words)P(words)}{P(sounds)} \tag{1}$$

- $P(sounds|words)$ can be estimated from training data,
- $P(words)$ is a priori probability of the words,
- $P(sounds)$ is a priori probability of the sounds.

During the training phase acoustic features (e.g., MFCCs) are extracted from the speech recordings. The corresponding manual transcriptions (words) are represented with their phonetic transcriptions. The acoustic models capture the mapping between acoustic features and the phonetic transcriptions of words. Since the acoustic features are extracted from certain speech recordings, the models trained, will generally underperform when used in different environmental conditions (e.g., noise, room acoustics, distance from microphone, type of microphone). Figure 2 illustrates the differences in the spectral representation of the same speech recording ("one, two, three") recorded using close range microphone and a microphone at 2 meters distance.

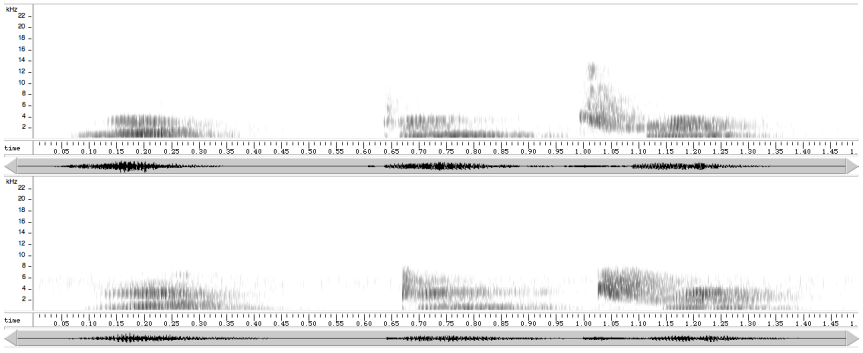

Figure 2: Microphone distance - close (top) and far (bottom)

One way to address the challenges arising from variability in environmental conditions is to adapt the acoustic models. This project presents one such scheme for adapting the acoustic models of an ASR system. The key idea is to transform the training data (speech recordings), as if it was recorded in the target environment. This involves, first to model the acoustics of the target environment and then, transform the original speech recordings to reflect the characteristics of that environment, and finally, train acoustic models on the transformed data.

2 Approach

For the purpose of testing the proposed scheme, we adapted speech recorded using close range microphones and used the transformed data to train acoustic models. In the remaining of this paper, we first introduce the steps to transform the original data and then, we present the results from performance evaluation.

Table 1 summarizes the differences between the original and the target environment used in this work.

Original environment		**Target environment**
Microphone	→	Kinect
Close distance	→	1 - 3 meters

Table 1: Characteristics of the original and the target environments

First step is to calculate the impulse response of the target microphone (Kinect) in the target environment and use it to transform the original data.

2.1 Impulse Response

To measure the impulse response we record a known signal with the target microphone (Kinect) in the target environment. Figure 3 shows the general recording setup.

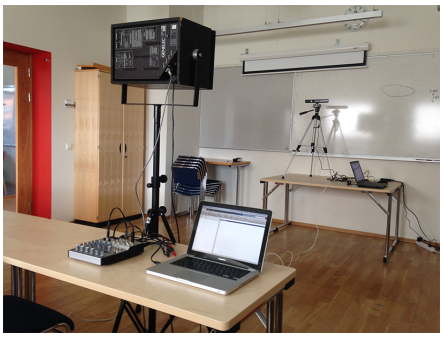

Figure 3: Speaker and microphone setup

In order to obtain better estimate of the impulse response, we record from 9 different speaker positions (see Figure 4).

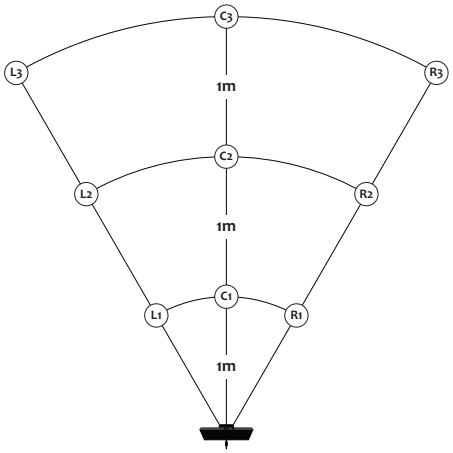

Figure 4: 9 positions from -60° to $+60^\circ$ with distances from 1m to 3m

Figure 5 illustrates the original and recorded signals. Observe that in the recorded signal, frequencies in the lower and the higher end are missing.

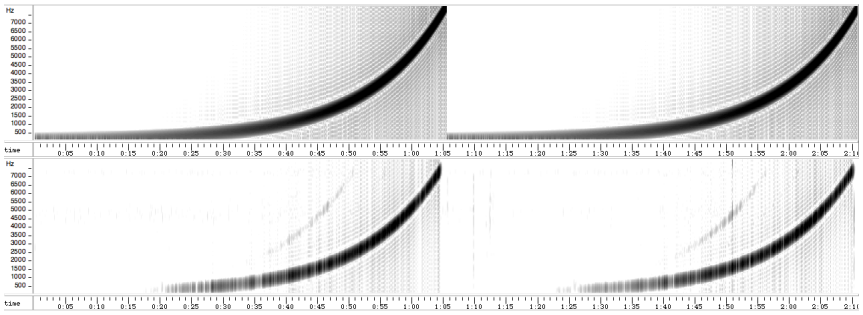

Figure 5: Original (top) and recorded (bottom) known signal

The impulse response is calculated by measuring by how much the frequencies in the original known signal get transformed into the frequencies of the recorded signal (see Equation 2).

$$h(t) = IFFT\left(\frac{FFT(\texttt{recorded signal})}{FFT(\texttt{known signal})}\right) \tag{2}$$

2.2 Data Transformation

Using the impulse response we transform the original speech recordings to reflect the microphone and the room acoustic characteristics. The transformation is done by scaling the frequencies in the original speech signal with the measured impulse response. This is obtained by solving Equation 3.

$$\texttt{transformed data} = IFFT\Big(FFT\big(\texttt{original data}\big) \times FFT\big(h(t)\big)\Big) \tag{3}$$

2.3 Models Training

Now that we have transformed the original speech recordings to reflect the characteristics of the target environment, we can train new acoustic models for the ASR system.

3 Evaluation

To evaluate the scheme presented here we compare the recognition performance of an ASR system using acoustic models trained on the original data with that of a system using acoustic models trained on the transformed data. For this we have used the HTK. We trained the system for recognizing Swedish sentences composed of 4 digits each. For training the system we used the TMHDIGIT dataset (2550[1] sentences). The data has been collected since 2012 as part of student lab exercises on speech recognition at TMH. For evaluating the recognizer's actual performance a separate test set was recorded in the target environment.

[1]We lost one sample on the way

One of the co-authors spoke 10 Swedish sentences (comprising of 4 digits each) at each of the 9 positions (see Figure 4). This resulted in a test set of 90 sentences.

In the following sections we present the results from various combination of training and test sets. For performance comparison we report the percentage of correctly recognized sentences (**SENT**), percentage of correctly recognized words (**WORD**), and word accuracy (**ACCU**).

3.1 Original Data

We trained the system on the original dataset and tested on test data for each of the 9 positions and also on the complete set (**Overall**). Table 2 summarizes the results of this test.

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{L_1}$	$\mathbf{L_2}$	$\mathbf{L_3}$	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_3}$	**Overall**
SENT	70.0	60.0	50.0	30.0	30.0	30.0	60.0	30.0	70.0	47.8
WORD	90.0	90.0	82.5	80.0	80.0	80.0	87.5	80.0	92.5	84.7
ACCU	90.0	90.0	80.0	75.0	77.5	77.5	87.5	77.5	90.0	82.8

Table 2: Performance (%) on the original training set (N = 2549)

3.2 Transformed Data

We transformed the original dataset for each of the 9 positions which resulted in a new dataset of $2549 \times 9 = 22941$ sentences. We trained the system on this transformed dataset and tested on test data for each of the 9 positions. Table 3 summarizes the results of this test.

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{L_1}$	$\mathbf{L_2}$	$\mathbf{L_3}$	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_3}$	**Overall**
SENT	70.0	40.0	70.0	50.0	60.0	70.0	60.0	20.0	40.0	53.3
WORD	92.5	97.5	90.0	90.0	95.0	92.5	92.5	87.5	92.5	92.2
ACCU	90.0	80.0	90.0	82.5	90.5	87.5	90.0	62.5	85.0	84.2

Table 3: Performance (%) on the transformed data (N = 22941)

Figure 6 provides a graphical overview of the trends in recognition performance (sentence level) at the 9 speaker positions using the transformed data in comparison to the original data. A black dot indicates improvement in recognition performance for that location, a gray dot suggests no improvement, and a white dot suggests decrease in recognition performance.

3.3 Transformed Data for $\mathbf{C_1}$

We trained the system on the transformed dataset corresponding to position $\mathbf{C_1}$, and tested on test data for each of the 9 positions.

	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{L_1}$	$\mathbf{L_2}$	$\mathbf{L_3}$	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_3}$	**Overall**
SENT	70.0	50.0	70.0	40.0	70.0	70.0	50.0	60.0	80.0	62.2
WORD	87.5	95.0	92.5	90.0	92.5	92.5	90.0	90.0	95.0	91.7
ACCU	87.5	85.0	92.5	85.0	92.5	90.0	87.5	82.5	95.0	88.6

Table 4: Performance (%) on the transformed data (N = 2549) for position $\mathbf{C_1}$

	C_1	C_2	C_3	L_1	L_2	L_3	R_1	R_2	R_3	Overall
SENT	60.0	30.0	60.0	40.0	50.0	80.0	60.0	20.0	30.0	47.8
WORD	90.0	90.0	92.0	87.5	92.5	95.0	92.5	90.0	92.5	91.4
ACCU	85.0	70.0	87.0	80.5	87.5	90.0	90.0	60.0	80.0	81.1

Table 5: Performance (%) on the transformed data (N = 2549) for position $\mathbf{C_2}$

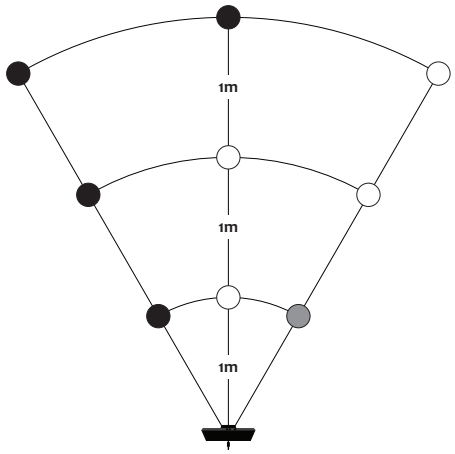

Figure 8: Trends in recognition performance for the 9 positions ($\mathbf{C_2}$)

3.5 Transformed Data for C_3

We trained the system on the transformed dataset corresponding to position $\mathbf{C_3}$, and tested on test data for each of the 9 positions.

	C_1	C_2	C_3	L_1	L_2	L_3	R_1	R_2	R_3	Overall
SENT	60.0	30.0	80.0	50.0	50.0	40.0	60.0	20.0	40.0	47.8
WORD	90.0	90.0	92.5	85.0	95.0	87.5	95.0	90.0	92.5	90.8
ACCU	87.5	75.0	92.5	77.5	87.5	75.0	90.0	60.0	82.5	80.8

Table 6: Performance (%) on the transformed data (N = 2549) for position $\mathbf{C_3}$

Figure 9 provides a graphical overview of the trends in recognition performance (sentence level) at the 9 speaker positions using the transformed data for $\mathbf{C_3}$ in comparison to the original data.

4 Conclusion

We have proposed a scheme for adapting the acoustic models of an ASR system to address the challenges arising from variability (noise, room acoustics, distance from microphone, type of microphone) in the recognition environment. The key idea in the proposed scheme is to transform the data (speech recordings) used for training the acoustic models for speech recognition, as if it was recorded in the target environment. This involves, first to model the acoustics of the target environment and then, transform the original speech recordings to reflect the characteristics of that environment, and finally, train acoustic models on the transformed data.

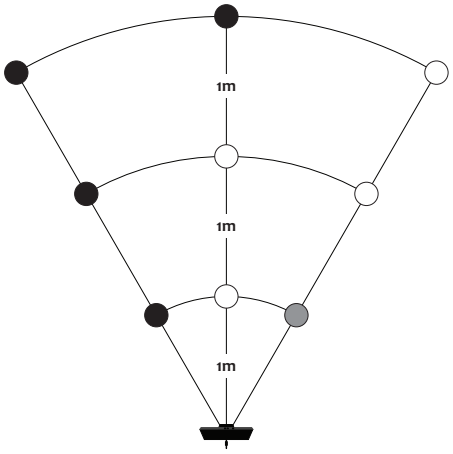

Figure 9: Trends in recognition performance for the 9 positions ($\mathbf{C_3}$)

We have evaluated the proposed scheme using HTK for speech recognition of Swedish sentences composed of 4 digits (continuous recognition). The results from the evaluation suggest that the overall sentence level accuracy improves from 47.8% to 53.3% when using acoustic models trained on transformed data (9 different positions). We have also presented the variations in recognition performance with respect to the variability in the transformed data (number of positions in the target environment). Interestingly, the recognition performance obtained using only the data adapted for a microphone distance of 1 meter (right in front of the speaker) offered a much higher overall performance - 62.2%. This may suggest that measuring the impulse response for a target environment, for one of the positions, may be sufficient to obtain a reasonable performance. These results are encouraging and suggest that the proposed scheme for acoustic adaptation has merits for addressing the challenge for speech recognition arising due to variability in the recognition environment.

A task for future work would be to do real time recognition in the target environment and evaluate the performance.

Acknowledgment

We would like to thank Giampiero Salvi for not only suggesting this topic for the project, but also for the many discussions involved at various stage of this work. The scheme for acoustic adaptation used here was originally proposed by Seshadri Sridharan (in Microphone, Acoustics Adaptation for Speech Recognition, Carnegie Mellon University Speech Group - 2013). We would like to thank them for sharing their ideas and Matlab implementations with us.