ACOUSTIC SYNTHESIS OF TRAINING DATA FOR SPEECH RECOGNITION IN LIVING ROOM ENVIRONMENTS

Volker Stahl, Alexander Fischer, Rolf Bippus

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
{Volker.Stahl, Alexander.Fischer, Rolf.Bippus}@philips.com

ABSTRACT

Despite continuous progress in robust automatic speech recognition in recent years acoustic mismatch between training and test conditions is still a major problem. Consequently, large speech collections must be conducted in many environments. An alternative approach is to generate training data synthetically by filtering clean speech with impulse responses and/or adding noise signals from the target domain. We compare the performance of a speech recognizer trained on recorded speech in the target domain with a system trained on suitably transformed clean speech. In order to obtain comparable results, our experiments are based on two channel recordings with a close talk and a distant microphone which produce the clean signal and the target domain signal respectively. By filtering and adding noise we obtain error rates which are only 10% higher for natural number recognition and 30% higher for a command recognition task compared to training with target domain data.

1. INTRODUCTION

A crucial factor for the performance of present speech recognition systems is that training and test conditions are acoustically similar. This means that training data collections must be conducted in every environment a recognition system is supposed to operate in, e.g. car, office, telephone, living room, etc. Even worse, such environments are usually parameterized, e.g. speed at which the car is driving, reverberation time of the living room, microphone distance, etc. As such data collections are both expensive and inflexible we investigate methods to generate training data synthetically by transforming clean speech under certain model assumptions of the target domain. The transformations we investigate are convolution with an impulse response and addition of a noise signal. The impulse response and the noise power spectrum have been measured in the target domain.

In order to evaluate this approach, time synchronous recordings with a high quality close talk microphone and an inexpensive distant microphone have been made in two living rooms. A speech recognition system is trained with various transformations of the close talk signal and tested on the distant microphone recording. The recognition accuracy is compared with a system trained on the distant microphone signal (matched scenario). Two recognition tasks are studied where the lexicon consists either of natural numbers or of command phrases.

The transformations of the clean signal improve the recognition accuracy significantly compared to a complete mismatch scenario and sometimes achieve the performance of matched training. Surprisingly, the influence of adding noise is more decisive than convolution with the target impulse response in our experiments. Further, as experiments with white noise indicate, the spectral shape of the noise has less influence than we expected.

The effect of various additive noise signals on speech recognition accuracy has been studied in [5]. Transforming clean speech by filtering, adding noise and MLLR adaptation for hands-free microphone array applications has been investigated in [1, 3, 4]. This paper is a follow-up of our previous work on transforming clean speech to the car environment by adding white noise and recorded car noise [2].

The concrete motivation for this work are large vocabulary conversational user interfaces for TVs, VCRs, audio sets, etc. Such applications must be speaker independent and require therefore large amounts of training material. Moreover, several languages need to be supported, which emphasizes even more the need for alternative methods to generate the training data.

The paper is structured as follows: Section 2 describes the recording conditions, speech data bases and the recognition system. In Section 3 we describe various ways how the clean signal was transformed into training data. We investigate convolutions with measured impulse responses, addition of white noise and colored noise whose spectrum corresponds to the target domain noise and combinations thereof. The resulting recognition error rates are reported in Section 4. Conclusions are drawn in Section 5.

2. EXPERIMENT SETUP

2.1. Recording Environment

The recordings were made in two different rooms which are 3m $\times$ 5m and 2m $\times$ 3m large and 2.5m high. Both rooms have a reverberation time T60 of about 0.44 seconds at 500 Hz. Two synchronous recordings with a near (0.4m) and a far (2.5m) microphone have been made and the impulse responses between speaker and the microphones have been measured, see Figure 1. The near microphone is a high quality AKG C1000 with an SPL Mike Man Model 9223 pre amplifier. The far microphone is a MWM MH118HC with a Radio Design Labs STM-1 pre amplifier. The MWM MH118HC is an inexpensive microphone of the kind which are used in consumer products. A sample utterance spoken in room 1 and recorded by both microphones is plotted in Figure 2. The impulse responses in room 1 are plotted for both microphone positions in Figure 3. For each utterance the SNR has been estimated using an automatic segmenter. The resulting SNR histograms for each microphone are plotted in Figure 4. As expected, the near microphone SNR is higher than the far microphone SNR. Finally the average noise power spectrum has been estimated for the far microphone in both rooms, see Figure 5.

The sampling rate was 32 kHz for the speech recordings and 44.1 kHz for the measurement of the impulse responses. All experiments were carried out on 8 kHz down sampled signals. The near and the far microphone signal are denoted by x_n and x_f, the corresponding impulse responses between speaker and microphones are denoted by h_n and h_f respectively.

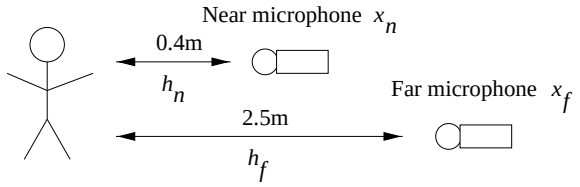

Fig. 1. Synchronous two microphone recording

2.2. Speech Database

The recorded database comprises 200 speakers (98 men, 102 women) with an age ranging between 10 and 60 years. The language is native British English. Tests were carried out on two sub corpora. The first corpus contains natural numbers, e.g. *"two thousand and three"*, *"seventeen"*, *"eighty"*, etc. and comprises 5288 utterances (3362 training, 1926 test, disjoint speakers). The lexicon consists of 32 words, which are modeled by whole word HMMs whose lengths range from 10 to 51 states. This is a difficult task because many words sound very similar, e.g. *"ninety"* and

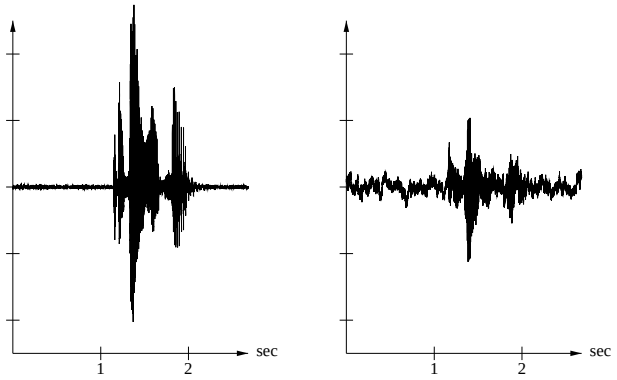

Fig. 2. Recording of the same utterance *"two thousand and four"* with near microphone (left graph) and far microphone (right graph) in room 1.

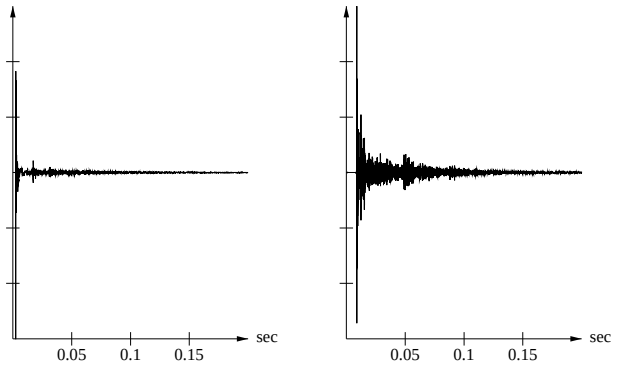

Fig. 3. First 200ms of impulse response from speaker position to near microphone (left graph) and far microphone (right graph) in room 1.

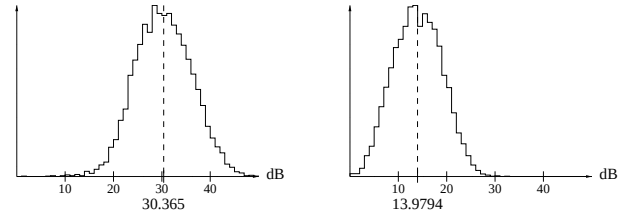

Fig. 4. Utterance based SNR histogram for near microphone (left graph) and far microphone (right graph) after preemphasis accumulated over both rooms.

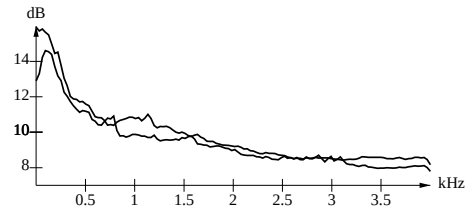

Fig. 5. Estimated logarithmic noise power spectrum from the far microphone in room 1 and room 2. The noise power in the near microphone is negligible.

"nineteen". The second corpus contains typical commands for consumer electronics devices, e.g. *"volume higher"*, *"silent"*, *"channel down"*, etc. and comprises 9844 utterances (6281 training, 3563 test). The lexicon consists of 54 words, which are modeled by whole word HMMs whose lengths range from 14 to 59 states.

2.3. Speech Recognition System

The speech recognizer is a continuous mixture density whole word HMM system whose parameters are estimated by Viterbi training. Each mixture consists of 8 Laplacian densities with a global diagonal covariance matrix. As our intention is to understand acoustic effects we refrained from using a language model or a grammar during recognition. The signal analysis is as follows: The observed speech signal is preemphasized and subdivided into overlapping, 16 ms spaced frames of 32 ms length. For each frame the power spectrum is estimated through a Hamming windowed FFT followed by a filter bank with 15 mel spaced triangular kernels. In order to reduce low frequency noise the filter bank cuts off frequencies below 187 Hz. The center frequency of the left most kernel is 312 Hz. After a discrete cosine transform of the logarithmic filter bank outputs 12 mel frequency cepstral coefficients are obtained whose long term mean is eliminated by a first order recursive filter with a large time constant. Finally, the feature vector is augmented by 12 coefficients obtained by linear regression over two frames in the past and two frames in the future. No further steps are taken for channel equalization or noise suppression.

3. INVESTIGATED APPROACHES TO SYNTHESIZE THE FAR MICROPHONE SIGNAL

In this section we define seven transformations which we applied to the near microphone signal x_n in order to generate training material:

- **(1)** x_n: Near microphone signal. This corresponds to a complete mismatch scenario, which means that the near microphone recording x_n is used for training and the far microphone recording x_f is used for testing.
- **(2)** $x_n * h_f$: Near microphone signal filtered with the transfer function from speaker to far microphone. The reverberation of the resulting signal is similar to x_f. The output signal has been scaled such that the speech energy in each utterance is the same as x_f. The measure we used for the speech energy of an utterance is the average energy in the 1500–3000Hz band of the 20% frames with highest energy. Perceptually **(2)** sounds like x_f except for additive noise.
- **(3)** $x_n * h_f * h_n^{-1}$: Like **(2)** but additional inverse filtering with the transfer function from speaker to near microphone. This is slightly more accurate than **(2)**, however we could not hear any difference between **(2)** and **(3)**.
- **(4)** $x_n * h_f * h_n^{-1} + n_c$: Like **(3)** but addition of a stationary noise signal n_c with the same power spectrum as the noise signal in x_f. The noise signal has been generated by estimating the average noise power spectrum of x_f over the entire training corpus, see Figure 5. Assuming zero phase, a complex noise spectrum has been generated which was used used to filter white noise. The result is a random signal with the same power spectrum as the average noise signal in x_f. Perceptually we did not realize any difference between **(4)** and x_f.
- **(5)** $x_n * h_f * h_n^{-1} + n_w$: Like **(4)** except that the added noise signal is white noise n_w with the same power as the noise in x_f. The difference in the background noise of **(5)** and x_f is clearly audible.
- **(6)** $x_n + n_c$: Assuming that the impulse responses are not given, the near microphone signal is simply added with a noise signal as in **(4)**. Before adding the noise x_n was scaled as described in $(\mathbf{2})$.
- **(7)** $x_n + n_w$: Like **(6)** except that white noise has been used.

The resulting signals are more or less similar to the far microphone signal x_f depending on the amount of information used from the target domain. Options **(2)** to **(5)** presume knowledge of impulse responses. Options **(4)** and **(6)** presume that the average noise power spectrum and the speech energy in x_f can be estimated. Option **(5)** and **(7)** merely require that the speech energy and average noise energy in x_f are given.

4. EXPERIMENTAL RESULTS

Table 1 and 2 summarize the speech recognition results for all combinations of corpora and rooms. Apart from error rates on the far microphone signal x_f we measured also the error rates when the test signals is generated by the same transformation of x_n as the training signal. The parameters of the speech recognizer have been optimized for the case when training and test is done on the target signal x_f.

According to our results, additive noise is more decisive than convolution with the target domain impulse response, even if the spectral shape of the noise is different from the target domain. The lowest error rates, which are about 10% higher for natural numbers and 30% for command phrases relative to training on x_f are obtained by combining filtering and adding noise. If x_n is used directly for training without any transformation, the error rate increases by 100% relative for natural numbers and 250% for command phrases.

Natural Numbers	room 1	room 2	room 1+2
Training and test on x_f			
x_f	**22.7**	**24.8**	**22.7**
Training on transformed x_n, test on x_f			
x_n	54.7	41.5	44.0
$x_n * h_f$	45.4	36.1	36.0
$x_n * h_n^{-1} * h_f$	45.5	34.9	36.1
$x_n * h_n^{-1} * h_f + n_c$	30.5	**22.5**	**24.7**
$x_n * h_n^{-1} * h_f + n_w$	**28.3**	30.2	27.3
$x_n + n_c$	42.6	26.2	28.6
$x_n + n_w$	37.2	34.4	32.4
Training and test on transformed x_n			
x_n	9.7	9.0	10.1
$x_n * h_f$	15.7	19.2	17.3
$x_n * h_n^{-1} * h_f$	16.2	17.8	17.5
$x_n * h_n^{-1} * h_f + n_c$	17.0	19.1	18.1
$x_n * h_n^{-1} * h_f + n_w$	21.6	24.6	25.2
$x_n + n_c$	9.2	11.3	10.5
$x_n + n_w$	14.5	20.3	17.3

Table 1. Word error rates on the natural number corpus. The first row is the baseline where the far microphone recording was used for training and testing. In the block below transformations of the near signal x_n are used for training and the far microphone recording x_f is used for testing. In the lower block training and test data are obtained by the same transformations of the near signal x_n.

Command phrases	room 1	room 2	room 1+2
Training and test on x_f			
x_f	**14.2**	**10.5**	**10.2**
Training on transformed x_n, test on x_f			
x_n	45.5	36.6	36.4
$x_n * h_f$	37.4	28.5	28.8
$x_n * h_n^{-1} * h_f$	35.3	26.9	27.7
$x_n * h_n^{-1} * h_f + n_c$	23.1	**13.7**	**14.7**
$x_n * h_n^{-1} * h_f + n_w$	**19.7**	16.5	16.0
$x_n + n_c$	33.2	18.9	20.9
$x_n + n_w$	30.1	19.5	19.8
Training and test on transformed x_n			
x_n	2.6	2.2	1.9
$x_n * h_f$	6.1	7.0	6.0
$x_n * h_n^{-1} * h_f$	7.0	7.0	6.0
$x_n * h_n^{-1} * h_f + n_c$	7.4	7.5	6.6
$x_n * h_n^{-1} * h_f + n_w$	9.6	12.8	10.3
$x_n + n_c$	3.0	3.0	2.4
$x_n + n_w$	5.1	7.3	5.3

Table 2. Same results as in Table 1 but for the command phrases corpus.

5. CONCLUSION

We compared the recognition accuracy of a system trained in the target domain with one trained on suitably transformed clean speech. Various transformations have been implemented which consist of convolution with measured impulse responses and adding white and colored noise. The experiments were carried out in two living rooms on two whole word recognition tasks. Best results were usually achieved by convolution with the impulse response between speaker and target microphone position, inverse convolution with the impulse response between speaker and close talk microphone position and addition of stationary noise whose power spectrum has been measured in the target domain. With this setup we obtain error rates which are 10% higher compared to matched training for natural numbers and 30% higher for command phrases.

The approach of using transformed clean speech for training has not only the advantage of reducing the effort for data collections. We hope to improve the recognition performance by using the filtered signal for Viterbi segmentation *before* the addition of noise. Further, we intend to use transformed training material with different instances of the noise process in order to obtain a more robust acoustic model.

6. REFERENCES

[1] M. Matassoni, M. Omologo and D. Giuliani, "*Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation*", in *Proc. ICASSP*, Vol. 2, pp. 1407–1410, June 2000

[2] R. Bippus, A. Fischer and V. Stahl, "*Domain Adaptation for Robust Automatic Speech Recognition in Car Environments.*", in *Proc. Eurospeech*, (Budapest, Hungary), pp. 1943–1947, September 1999

[3] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "*Training of HMM with filtered speech material for hands-free speech recognition*", in *Proc. ICASSP*, Vol. 1, pp. 449–452, March 1999

[4] M. Omologo, O. Svaizer and M. Matassoni "*Environmental conditions and acoustic transduction in hands-free speech recognition*", Speech Communication, Vol. 25, pp. 75–95, 1998

[5] A. Varga, H. J. M. Steeneken, M. Tomlinson and J. Jones, "*The NOISEX-92 study on the effect of additive noise on automatic speech recognition*", Booklet included in the NOISEX-92 CD-ROM set.

[6] S. Haykin "*Adaptive Filter Theory.*", Englewood Cliffs, NJ (USA), Prentice-Hall, 1991