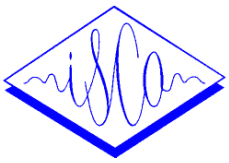

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A FEATURE-SPACE TRANSFORMATION FOR TELEPHONE BASED SPEECH RECOGNITION

Alexandros Potamianos † *Li Lee* † *and Richard C. Rose*

AT&T Bell Labs, Murray Hill, NJ 07974, U.S.A.

ABSTRACT

An experimental study describing the effects of carbon and electret telephone transducers on automatic speech recognition (ASR) performance is presented. It is shown that telephone based ASR performance on a connected digit task actually improves when speech is spoken through the carbon transducer. This surprising result is explained by a study of the differences in acoustic characteristics between carbon and electret telephone handsets. An initial attempt is made to devise a simple procedure for obtaining a parametric transformation which emulates the properties of the carbon transducer. The parameters of this transformation are trained automatically from speech spoken simultaneously through carbon and electret telephone handsets. When telephone speech data is transformed according to this procedure, a significant improvement in ASR performance is obtained. These results are interpreted and future research directions are discussed.

1. INTRODUCTION

Despite the fact that a large percentage of telephones in most telecommunications markets employ carbon transducers, the physical and acoustic characteristics of the carbon transducer are poorly understood. It is known that carbon transducers are highly non–linear, with characteristics that vary considerably over time and from one transducer to the next [1]. Yet, despite the nonlinear distortion and variability that the carbon button transducer introduces to telephone speech, the carbon has long been known to be preferred by human listeners over transducers with more linear characteristics. Formal subjective studies have shown that human listeners prefer speech spoken through carbon telephone handsets over speech spoken through electret handsets [2]. As a result, expander circuitry is included in the design of some telephone handset models containing electret transducers to emulate the background noise reduction properties of carbon transducers [2].

In this paper, we present evidence that the acoustic characteristics of carbon transducers can actually enhance automatic speech recognition (ASR) performance. Specifically, it is shown that performing ASR on utterances spoken through a carbon transducer results in improved performance over that obtained using an electret transducer. A 50% reduction in ASR error rate was observed for utterances spoken through carbon transducers for a connected digit recognition task over the public switched telephone network.

In section 3, empirical evidence is provided relating to properties of the carbon transducer that explain this surprising result. We argue that it is the carbon transducer's relative insensitivity to turbulent airflow phenomena which makes it less sensitive to many of the sources of variability that can be detrimental to ASR performance. We also present evidence which suggests that many of the non–linear distortions commonly attributed to the carbon microphone may have a less significant effect on the feature representations that are used in ASR than turbulent airflow phenomena. Finally, we incorporate the acquired knowledge of the properties of the carbon transducer to derive a nonlinear feature space transformation to emulate the carbon transducer's desirable characteristics. Our ultimate goal is to improve speech recognition performance over the public switched telephone network (PSTN).

2. EFFECT OF TRANSDUCER TYPE ON ASR PERFORMANCE

An experimental study was performed to determine the effects of using carbon as opposed to electret based telephone handsets on ASR performance. The speech corpus used for the study was collected over the PSTN. The utterances collected were spoken through both carbon and electret transducers, allowing HMM training and testing to be performed on carbon or electret utterances separately. The training/testing data consisted of 5368/2239 continuous utterances spoken over the telephone network by 52/22 speakers. Each utterance contained from one to seven digits with the total number of digits in the training/testing data amounting to 16321/6793.

Speech recognition was performed using continuous Gaussian observation density hidden Markov models (HMM) defined over mel-frequency cepstrum observations, their derivatives, and their second derivatives. It is important to note that the speech data collection was performed under controlled circumstances. All utterances were collected using a set of two electret and two carbon telephone handsets which were tested and found to be in good working condition. Previous studies of ASR performance over the PSTN relied on the use of telephone station sets that already existed in the speakers' homes. These station sets, especially those based on carbon transducers, can often be in poor working condition.

The per digit error rates for different training and testing conditions are given in Table 1. There are several observations that can be made from this table. First, the error rate was lower for test utterances spoken through a carbon handset regardless of the training conditions. Second, for matched training and testing conditions, the performance for carbon handsets was almost 50% better than for electret handsets. Finally, it was observed that, when testing using the carbon handset utterances, the difference in performance between matched and unmatched conditions was relatively small. Obviously, any system that provides good overall ASR performance and robust-

† Alexandros Potamianos is currently with School of ECE, Georgia Institute of Technology, Atlanta, GA

† Li Lee is currently with Dept. of EECS, Massachusetts Institute of Technology, Cambridge, MA

10.21437/Eurospeech.1995-375

ness to varying training conditions is very desirable. A better understanding of the characteristics of the carbon transducer may help us in designing speech recognition systems with these properties.

Testing / HMM Training	Carbon Data	Electret Data
Carbon Data	1.3%	4.1%
Electret Data	1.6%	2.4%
Combined Data	1.3%	2.8%

Table 1: Telephone network based connected digit recognition error rate for transducer-dependent models.

3. ACOUSTIC PROPERTIES OF CARBON AND ELECTRET TRANSDUCERS

This section attempts to compare and quantify the differences in acoustic characteristics between the carbon and electret telephone transducers as they relate to automatic speech recognition. We look specifically at how the different transducers affect the underlying feature representations that are used in ASR.

In order to make direct comparisons between carbon and electret transducers and their effect on speech signals, a stereo database of carbon/electret speech (SCE) was used. Connected digit utterances were spoken simultaneously through carbon and electret handsets, and recorded over the PSTN. Thirty-one three-digit strings were recorded in two sessions for each of four different speakers, resulting in a total of 248 utterances. Initially, these recordings were used to make comparisons between the mel-frequency log filter-bank energies derived using the two transducers. A Davis and Mermelstein filter-bank consisting of twenty-four filters over a four kHz. bandwidth was used in all experiments. These comparisons are important because the cepstrum observation vectors used in the speech recognition experiments were derived from the filter-bank energies through an inverse cosine transform.

Fig. 1(a) displays the short-time energy contour for the 3-digit string 'three-six-six'. Note the differences in the average energy for the plosive /ks/ and the fricatives /s/, /th/ between the carbon transducer (solid line) and electret transducer (dashed line). The electret transducer is clearly affected by the direct airflow that accompanies plosive and fricative production. This 'pop' effect is well known and compensated for in some electret transducers through careful design of the microphone grill [3]. However, in most telephone handsets the principle mechanism for reducing the effects of the turbulent airflow is to design the handset so that the transducer is positioned outside the air stream. As a result, there is considerable variation in the amount of 'pop' noise depending on the positioning of the handset relative to the speakers' mouth. Indeed, we have observed different degrees of distortion among speakers and sessions in the SCE database. Average energy differences between carbon and electret utterances can be as high as 10 dB.

Fig. 1 (b) displays typical mel-frequency filter-bank envelopes measured in dB for the phonemes /iy/, /ah/, and /s/ from the SCE database. The spectral energy contours for carbon utterances (solid line) show spectral minima that are consistently attenuated relative to that of electret utterances (dashed line). The local carbon-electret spectral envelope differences can be as high as 20 dB in the vicinity of spectral minima. This is thought to

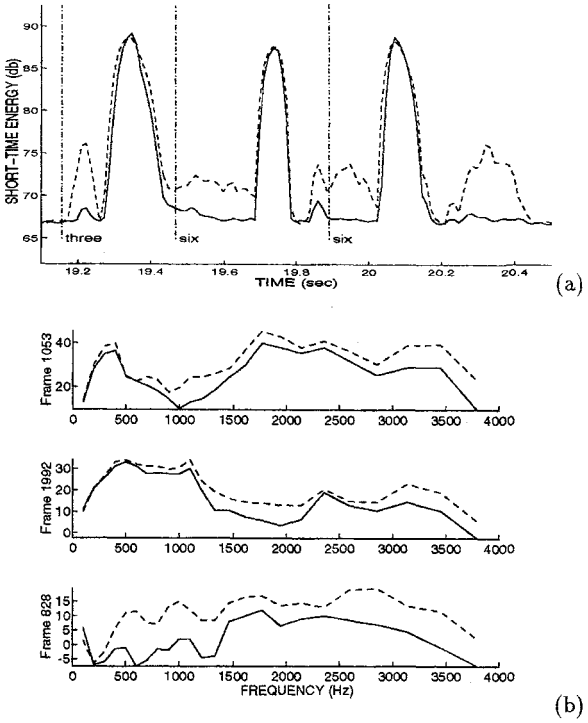

Figure 1: (a) Short-time energy contours for stereo carbon (solid) and electret (dashed) and carbon data for the utterance 'three-six-six' (b) Mel spectral envelope for stereo carbon (solid) and electret data speech frames (20 msec): /iy/ from 'three', /ah/ from 'one', /s/ from 'six'

be primarily a result of the turbulent airflow 'pop' effect that introduces noise in the spectrum. From the anecdotal evidence illustrated by the plots in Fig. 1, it is clear that there is considerably more variability associated with speech that is passed through the electret transducer. It is very likely that this increased variability is responsible for the increase in ASR error rates obtained using electret data.

4. NON-LINEAR CHARACTERISTICS OF THE CARBON TRANSDUCER

The discussion in the previous section suggested that the most salient difference between carbon and electret telephone handsets from an ASR point of view is the carbon transducer's insensitivity to turbulent airflow phenomena. There was no mention of the widely know non-linear properties associated with the carbon transducer. This section briefly discusses these issues and considers their relative importance for ASR.

The well established facts about the acoustic characteristics of the carbon transducer are [1]: (a) random variation of the sensitivity of the transducer depending on the state of agitation of the carbon granules which may include 'packing' or total loss of sensitivity when the carbon granules are packed together; (b) non-linear instantaneous distortion evident as clipping for 'expansive' signal peaks (i.e., peaks corresponding to carbon chamber expansion) and smooth compression of 'compressive' peaks; (c) nonlinear input/output gain curve resulting in suppression of background noise; (d) contact noise and noise surges caused by current instabilities in the network.

A laboratory study was performed to investigate the

properties of the carbon transducer by attaching a carbon telephone handset powered from the telephone network to a B&K head and torso simulator. The important difference between measurements obtained from this experimental set–up and the measurements discussed in Section 3 is the measurements obtained here were not taken from actual speakers. Therefore, the airflow effects should *not* be a factor. The carbon transducer's response to a number of different stimuli were analyzed including synthetic stimuli and a pre-recorded utterances. These were part of a larger set of experiments that were intended to extend a previous study of carbon transducer characteristics performed by Joe Hall and Bob Kubli using the same experimental apparatus [4].

In Fig. 2(a), we show the response of the carbon transducer to a pair of tones at 100 dB SPL. Intermodulation harmonics appear at multiples, sums, and differences of the original frequencies due to the nonlinear distortion introduced by the transducer. The fact that the amplitudes of the non–linearly generated harmonic tones are sometimes as little as five dB down from the input amplitudes illustrates the effect of the non–linear transducer on simple tones. Of course, the degree of distortion is heavily dependent on the input dynamic level.

In Fig. 2(b) we compare the mel-frequency filter–bank energies for the carbon transducer (solid line) and a reference condenser microphone (dashed line) for the phoneme /ae/ recorded at 100 dB SPL. The spectrum of the carbon transducer has been normalized by the response to a swept tone. For speech signals, one might expect that intermodulation products might severely distort the spectral envelope. However, comparisons of many curves like the ones in Fig. 2(b) suggest that, at least for the smoothed spectral representations that are obtained from the mel–frequency filterbank analysis, this effect is minor. The combined data described in Sections 3 and 4 suggest that the effect of the well known non–linearities in the carbon transducer may have far less of an effect on ASR performance than the turbulent airflow phenomena described in Section 3.

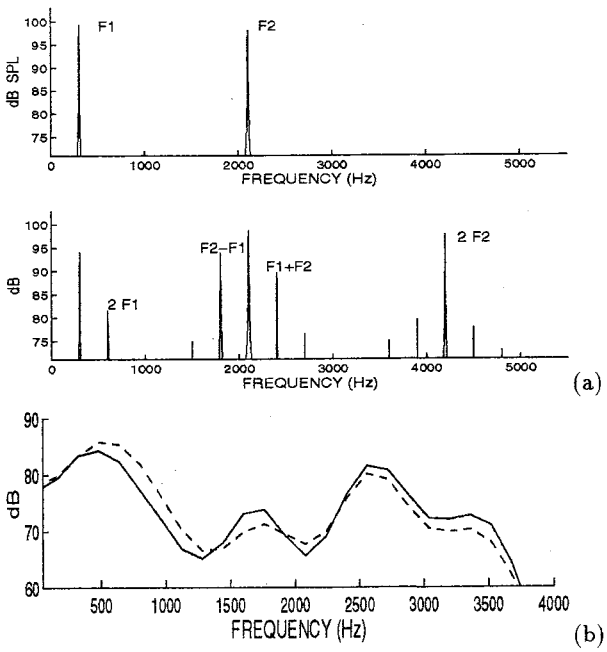

Figure 2: (a) Fourier spectrum for a pair of tones and for their carbon transducer output (b) Mel–frequency filter–bank spectrum for carbon (solid) v. reference (dashed) transducer output for /ae/ from 'vans'.

5. FEATURE SPACE TRANSFORMATION

An initial effort was made to model the differences in acoustic characteristics between the carbon and electret transducers. This model has been used in a procedure to reduce the transducer induced variability described in Table 1. The model was motivated by the experimental evidence suggesting that the desirable effects of the carbon transducer stem from its ability to suppress highly variable speech information. Hence, the general notion behind the compensation or transformation procedure is to try and recover the characteristics the carbon transducer from the utterances spoken through a "less desirable" linear transducer. Since the effects of this non–linear device are dependent on the characteristics of the input speech, it is important that the transformation be time–varying. The transformation corresponds to a set of HMM state dependent linear transformations, effectively approximating the non–linear characteristic as a segmental linear models.

The problem is treated as a signal recovery problem. The cepstrum vectors derived from the carbon transducer, $\vec{x}_t^c$, are taken as the "desired" signal, and the cepstrum vectors derived from the electret transducer, $\vec{x}_t^e$, are taken as the "corrupted" signal. It is assumed that these vectors are realizations of random processes that are related according to $\vec{x}_t^e = \vec{x}_t^c + \vec{y}_t$, where $\vec{y}_t$ represents a simple linear filtering operation. This can be modeled as an additive bias in the mel-frequency cepstrum domain. It is also assumed that $\vec{x}_t^c$ and $\vec{y}_t$ are both represented by Gaussian densities that are tied to the states of the hidden Markov digit models. The parameters of the HMM state dependent Gaussian densities associated with $\vec{y}_t$ are obtained from the simultaneous carbon / electret recordings of the SCE database according to the process illustrated in Fig. 3. Note that the set of speakers and transducers used in the SCE corpus are completely separate from those used in the corpus described in Section 2 which is used to evaluate recognition performance. Viterbi alignment of each training utterance spoken through a carbon transducer is performed against the known word transcription for the utterance. All frames where $\vec{x}_t^c$ are assigned to state $\theta_t = k$ are used to estimate the mean, $\vec{\mu}_k$, and variance, $\vec{\sigma}_k^2$ of $\vec{y}_t = \vec{x}_t^e - \vec{x}_t^c$ for state k. After obtaining these statistics as part of a training procedure, it is possible during recognition to estimate the expected value of the carbon transducer output conditioned on the electret transducer output for each state $E\{\vec{x}_t^c | \vec{x}_t^e, \theta_t = k\}$. However, in our experiments, using this optimal linear estimator performed no better than simply subtracting the mean $\vec{\mu}_k$ from observations decoded during recognition to state $\theta_t = k$.

During speech recognition, the compensation procedure is applied as shown in Figure 4. First, a list of N most likely string candidates (n–best list) is generated from the original test utterance. Second, a state dependent transformation is performed for each string candidate, by replacing each observation $\vec{x}_t^e$ with $\vec{x}_t^e - \vec{\mu}_{\theta_t}$ as described above. Finally, the best string is chosen as the one associated with the transformed utterance with the highest likelihood. This process of "rescoring" the top scoring string candidates produced by the Viterbi decoder is an easy way of applying acquired knowledge of transducer distortion. However, there is no reason why this procedure could not be more tightly coupled in the initial search procedure. The speech recognition performance obtained

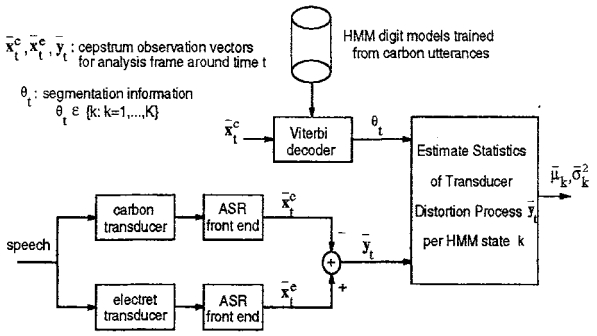

Figure 3: Characterizing the carbon–electret transducer distortion process. For each HMM state $\theta_t = k$, the mean $\vec{\mu}_k$ and variance $\vec{\sigma}_k^2$ of the "transducer distortion" cepstrum vectors $\vec{y}_t$ are computed. The state segmentation $\theta_1, \ldots, \theta_T$ of the input utterance is obtained for a known transcription using the carbon HMM.

using this procedure is measured for test utterances spoken through both electret and carbon transducers, as it is important that the error rate does not increase under any condition.

Table 2 displays the speaker independent digit recognition error rate using a number of different testing conditions. The results are reported for the same speech recognition corpus that is described in Section 2. In all cases, a hidden Markov model trained from combined data using both carbon and electret transducers is used for testing. The error rate displayed in the first row of Table 2, given as the baseline condition, is simply repeated from Table 1. The second row of Table 2 is the error rate for the compensation procedure described in Figure 4 when an n-best list consisting of the 4 best string candidates is used. The "n–best / CMS rescoring" procedure shown in the third row of the table is identical to the procedure shown in the second row except the compensation and rescoring is performed using cepstrum mean normalized (CMS) training and test data. Finally, the last row of the table represents the error rate obtained using cepstrum mean subtraction (CMS) on both training and test data.

Test Conditions	Word Error Rate	
	Carbon	Electret
Baseline	1.3%	2.8%
N-Best / Rescore	1.2%	2.4%
N-Best / CMS Rescore	1.0%	1.9%
CMS	1.1%	2.1%

Table 2: Connected digit recognition error rate for the transformed electret and carbon PSTN test data.

6. SUMMARY

The compensation procedure described in this paper was motivated by the observation that performing speech recognition on utterances spoken through carbon transducers can reduce error rate by over 50% compared to the performance obtained using electret transducers. There are several observations that can be made from the results in Table 2. The first observation is that the compensation procedure described in Section 5 improved performance for utterances obtained from electret transducers by over

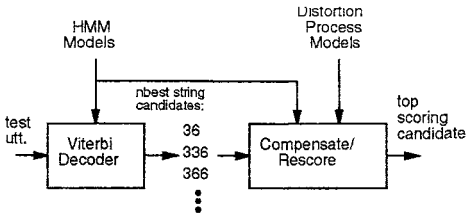

Figure 4: Applying transducer compensation process as part of an n–best string rescoring procedure.

30%. However, it is also true that simply performing cepstrum mean normalization (being careful to estimate cepstrum averages using only observation vectors corresponding speech frames) improved performance by 25%. Hence, while the compensation procedure was effective, there is still much to be done to impose a greater consistency between the process of parameter identification and recognition.

In general, it is very difficult to estimate the parameters of a spectral mapping procedure from the small number of observations that are available during recognition. The procedure described here attempts to deal with this problem by estimating these parameters *a priori*. Non–parametric probabilistic spectral mappings have been proposed in the past for speech restoration [5]. The simple parametric spectral mapping procedure investigated in this paper is important because if incorporates prior information obtained from acoustic observations from the non–linear carbon transducer.

7. ACKNOWLEDGEMENTS

We wish to thank Bob Kubli for his helpful advice and his invaluable assistance in setting up and calibrating the transducer measurement apparatus. We also thank Joe Hall for his advice and discussion concerning his own analysis of carbon transducer characteristics. We would also like to thank our colleagues Jim West, Gary Elko, and Cecil Coker for their many helpful comments. Finally, we thank David Roe for his role in the collection of the SCE speech database, and Richard Sachs for his role in the collection of the larger telephone digit database used in this study.

8. REFERENCES

[1] L. S. Moye, "Study of the effects on Speech Analysis of the Types of Degradation Occurring in Telephony," Tech. Report, Standard Telecommunication Laboratories Limited, 1979.

[2] R. L. Miller, personal communication, AT&T Bell Labs, 1994.

[3] M. L. Gayford et al, *Microphone engineering handbook*, Oxford; Boston: Focal Press, 1994.

[4] J. Hall, unpublished, AT&T Bell Labs, 1993.

[5] B.–H. Juang and L. R. Rabiner, "Signal Restoration by Spectral Mapping," in *Proc. Int. Conf. Acoustics, Speech, Signal Proc. '87*, pp. 2368-2371, Dallas, Texas, 1987.