WTIMIT: The TIMIT Speech Corpus Transmitted Over the 3G AMR Wideband Mobile Network

Patrick Bauer, David Scheler, Tim Fingscheidt

Technische Universität Braunschweig, Institute for Communications Technology,
Schleinitzstr. 22, D – 38106 Braunschweig, Germany
{bauer,scheler,fingscheidt}@ifn.ing.tu-bs.de

Abstract

In anticipation of upcoming mobile telephony services with higher speech quality, a wideband (50 Hz to 7 kHz) mobile telephony derivative of TIMIT has been recorded called *WTIMIT*. It opens up various scientific investigations; e.g., on speech quality and intelligibility, as well as on wideband upgrades of network-side interactive voice response (IVR) systems with retrained or bandwidth-extended acoustic models for automatic speech recognition (ASR). Wideband telephony could enable network-side speech recognition applications such as remote dictation or spelling without the need of distributed speech recognition techniques. The WTIMIT corpus was transmitted via two prepared Nokia 6220 mobile phones over T-Mobile's 3G wideband mobile network in The Hague, The Netherlands, employing the Adaptive Multirate Wideband (AMR-WB) speech codec. The paper presents observations of transmission effects and phoneme recognition experiments. It turns out that in the case of wideband telephony, server-side ASR should *not* be carried out by simply decimating received signals to 8 kHz and applying existent narrowband acoustic models. Nor do we recommend just simulating the AMR-WB codec for training of wideband acoustic models. Instead, real-world wideband telephony channel data (such as WTIMIT) provides the best training material for wideband IVR systems.

1. Introduction

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993) is intended for developing and evaluating automatic speech recognition (ASR) systems. It comprises wideband speech recordings (sampled at 16 kHz and containing frequency components from about 50 Hz to 7 kHz) of 630 native speakers of eight major dialect regions of the United States. Each of the speakers produced ten phonetically rich sentences. For each utterance, a speech waveform as well as time-aligned orthographic, phonetic, and word transcriptions are provided. Today, five TIMIT derivatives are already available: FFMTIMIT, NTIMIT, CTIMIT, HTIMIT, and STC-TIMIT. The FFMTIMIT (Free-Field Microphone TIMIT) corpus (Garofolo et al., 1996) consists of the original TIMIT database recorded with a free-field microphone. NTIMIT (Network TIMIT) serves as a telephone bandwidth adjunct to TIMIT (Jankowski et al., 1990). It contains the original TIMIT speech files transmitted over a telephone handset and the NYNEX telephone network, subject to a large variety of channel conditions. For the cellular bandwidth speech corpus CTIMIT (Brown and George, 1995), the original TIMIT recordings were passed through cellular telephone circuits. The HTIMIT (Handset TIMIT) corpus (Reynolds, 1997) offers a TIMIT subset of 192 male and 192 female speakers transmitted through different telephone handsets for the study of telephone transducer effects on speech. For the single-channel telephone corpus STC-TIMIT (Morales et al., 2008), the TIMIT recordings were sent through a real and, in contrast to NTIMIT, single telephone channel.

While some of these derivative TIMIT corpora are wideband speech, others are telephony corpora containing narrowband speech, i. e., sampled at 8 kHz and containing frequency components from about 300 Hz to 3.4 kHz. Until now, no *real-world wideband telephony* speech corpus is publicly available. Thanks to upcoming wideband speech codecs – historically starting from the ITU-T subband ADPCM speech codec G.722 in 1988 (ITU, 1988), via the ITU-T transform domain speech codec G.722.1 for low frame-loss in 1999 (ITU, 1999), the 3GPP / ITU-T Adaptive Multirate Wideband (AMR-WB) speech codec G.722.2 in 2001 (3GPP, 2001; ITU, 2002), to the recent G.711.1 wideband embedded extension for G.711 pulse code modulation in 2008 (ITU, 2008) – wideband telephony speech transmission is already feasible nowadays, even in an increasing number of mobile networks. Hence, a wideband telephone-bandwidth adjunct to TIMIT is desirable for the development and evaluation of systems and allows a wide range of scientific investigations; e. g., on possible wideband upgrades of network-side interactive voice response (IVR) systems with retrained or bandwidth-extended acoustic models for ASR (Bauer and Fingscheidt, 2009a). Over a wideband speech telephony system, network-side speech recognition could allow even advanced applications, such as remote dictation or spelling, which in the past could be deployed in mobile telephony services only under the distributed speech recognition paradigm. Further research topics could be investigations on the quality and intelligibility of wideband speech compared to narrowband speech as well as the influence on human cognitive load. Depending on the results, this could encourage legislation to make wideband handsfree equipment mandatory in cars, wherever wideband mobile networks are available.

Our paper presents a new corpus called WTIMIT (Wideband Mobile TIMIT) containing the recordings of the original TIMIT speech files after transmission over a 3rd generation (3G) AMR-WB mobile network. In the following, the most important steps for the generation of WTIMIT are detailed, such as data preparation, speech transmission and

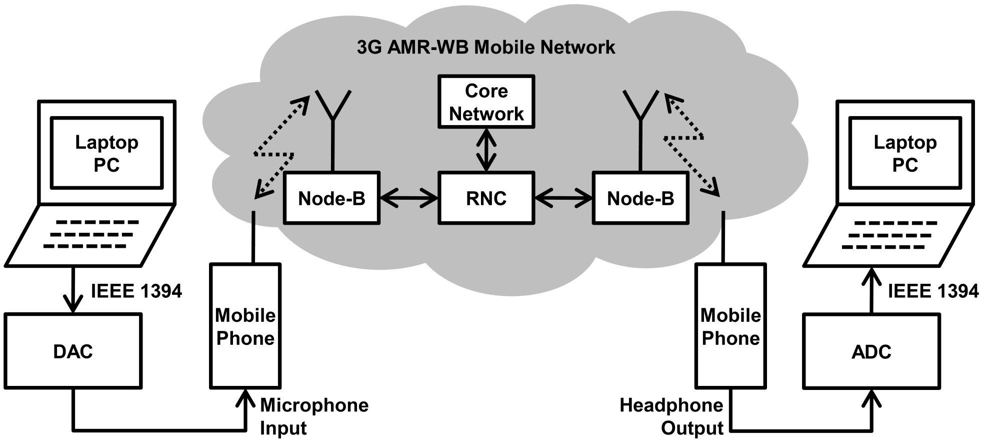

Figure 1: Setup of the TIMIT speech corpus transmission over a 3G wideband mobile network.

recording, as well as post-processing. The transmission effects are analyzed by means of calibration tones. Additionally, phoneme recognition results are provided based on WTIMIT as well as on TIMIT, and compared with those of other derivatives of TIMIT, such as a TIMIT version transmitted over a 3G narrowband mobile network. Finally, conclusions are drawn and the distribution of WTIMIT via the Linguistic Data Consortium (LDC) is explained.

2. Generation of WTIMIT

2.1. Data Preparation

All original TIMIT speech files dedicated to training and testing were first converted into raw PCM waveform data by dropping the first 1024 bytes of TIMIT's header information. The raw speech data was then concatenated to form 11 signal chunks of at most 30 min duration. In order to allow precise de-concatenation after transmission and to be able to examine the frequency-dependent codec influence and channel distortion (Section 3), each signal chunk was preceded by a 4 s calibration tone. It comprised 2 s of a 1 kHz sine wave followed by another 2 s of a linear sweep from 0 to 8 kHz.

2.2. Speech Transmission and Recording

The corpus transmission was conducted in T-Mobile's AMR-WB-capable 3G mobile network in The Hague, The Netherlands, according to the setup depicted in Figure 1.

At the sending end, the speech chunks were played back by a laptop PC. Via an IEEE 1394 link (FireWire), the data was transmitted digitally to an external digital-to-analog converter (DAC) of type RME Fireface 400. The analog signal was then fed electrically into the microphone input of the transmitting Nokia 6220 mobile phone. For this purpose, an audio quality test cable for Nokia mobile phones was used. Note that thanks to the electrical interface to the phone it was ensured that the common Nokia 6220 microphone equalization was bypassed.

Prior to the actual transmission, the output attenuation of the DAC was adjusted such as to prevent analog saturation at the input circuit of the phone while ensuring an optimal dynamic range. Furthermore, a call to the phone at the receiving end, a second mobile phone of type Nokia 6220, was established for each speech chunk separately. Using the field test monitoring software of the phones, we confirmed that they were situated in different network cells at all times during transmission; moreover, we verified that the proper speech codec – the widely used AMR-WB (also known as G.722.2) – was being employed at a constant bitrate of 12.65 kbit/s. Note that this bitrate is by far the most widely used one.

At the receiving end, the analog headphone output of the receiving mobile phone was connected electrically to an analog-to-digital converter (ADC) of type RME Fireface 400. The analog input gain of the latter device was adjusted once initially to exploit the dynamic range of the ADC. Sampling was performed at a rate of 48 kHz, the native sampling rate of the ADC, and with 16 bit precision. The digital speech signals were transferred to a laptop PC again via an IEEE 1394 link and recorded onto a hard drive.

2.3. Post-Processing

The transmitted speech chunks were decimated from 48 kHz to 16 kHz sampling rate using a high-quality lowpass filter. Finally, they were de-concatenated by maximizing the normalized cross-correlation between them and the original speech files. We followed the deconcatenation methodology of STC-TIMIT as described in (Morales et al., 2008), in order to obtain a precise sample alignment to the TIMIT speech files. Due to jitter effects of the packet-switched 3G network, parts of some WTIMIT speech files are slightly misaligned against the original time alignments (Section 3). This may especially impact the location of short, non-stationary phonemes and speech pauses. However, TIMIT's original label files can still be expected to largely apply to WTIMIT.

3. Observations of Transmission Effects

The frequency-dependent codec influence and channel distortion of the transmission are examined by means of calibration tones. Figure 2 depicts the spectrograms and waveforms of the original calibration tone (left) and of a transmitted example (right). Clearly observable effects of the transmission are some noise, and the upper cut-off frequency of the linear sweep at about 6 kHz. Note further

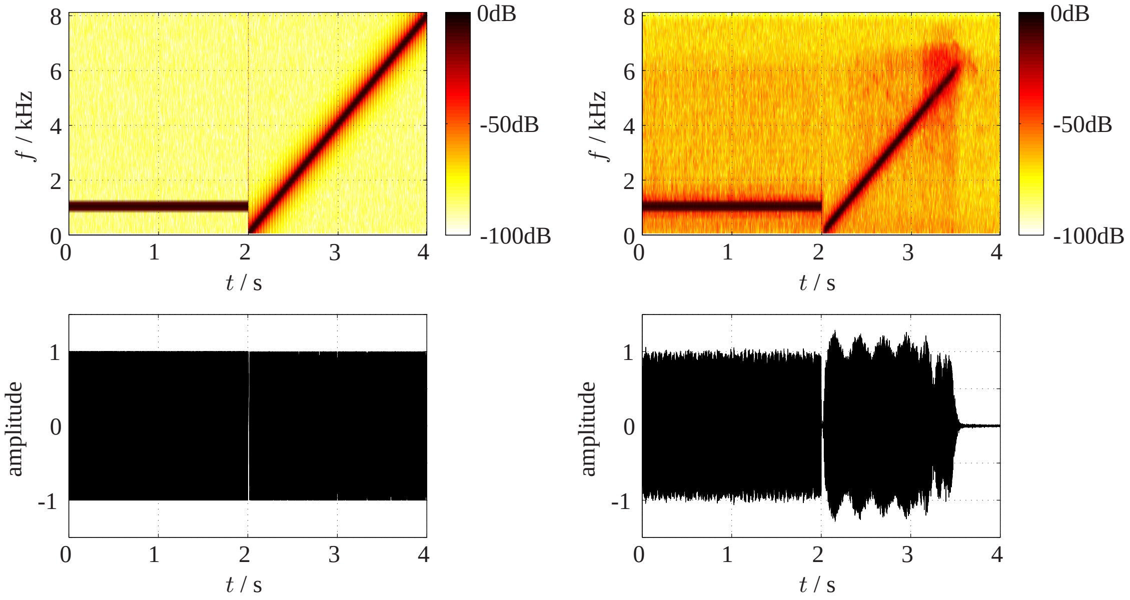

Figure 2: Spectrogram and waveform of two calibration tones: Original one (left) vs. exemplarily transmitted one (right).

in the waveform of the transmitted calibration tone the rippled amplitude characteristic of the linear sweep.

Now and then, slight misalignments of about 10 to 20 ms between the original TIMIT and the transmitted WTIMIT speech files were found to be produced by the channel mainly during speech pauses. These channel effects are related to the packet switching domain in the 3G Core Network. Depending on the traffic load in the network, packets are buffered and queued, which results in a variable packet delay, that is jitter.

In rare cases, stronger noise artifacts could be observed, like a sudden increase of the sound level or short bursts of disturbing sinusoidal signals. Sometimes, speech parts were also truncated. Please note that these effects may have arisen due to the pilot character of the mobile network in the experimentally operating AMR-WB mode. Anyhow, the overall perceived quality and particularly the intelligibility of the speech transmission was informally judged to be excellent.

4. Phoneme Recognition Experiments

4.1. Experimental Setup

For an objective assessment of the bandwidth, codec, and channel effects of WTIMIT we conducted phoneme recognition experiments. For comparison, the same experiments were conducted on several corpora derived from TIMIT:

- TIMIT:
  Original TIMIT corpus.

- TIMIT–AMRWB1265(sim):
  TIMIT corpus subject to AMR-WB fixed-point codec simulation at bitrate 12.65 kbit/s.

- WTIMIT = TIMIT–AMRWB1265(3G):
  TIMIT corpus transmitted over a real 3G AMR-WB mobile network at bitrate 12.65 kbit/s.

- TIMIT↓2:
  TIMIT corpus decimated to 8 kHz sampling rate using a high-quality lowpass filter.

- TIMIT↓2–AMRNB122(sim):
  Decimated TIMIT corpus subject to AMR-NB fixed-point codec simulation at bitrate 12.2 kbit/s.

- TIMIT↓2–AMRNB122(3G):
  Decimated TIMIT corpus transmitted over a real 3G AMR-NB mobile network mainly at bitrate 12.2 kbit/s.

- WTIMIT↓2:
  WTIMIT corpus decimated to 8 kHz sampling rate using a high-quality lowpass filter.

Note that the derivative corpus TIMIT↓2–AMRNB122(3G) was created in analogy to WTIMIT. However, the transmission was conducted over T-Mobile’s 3G mobile network in Braunschweig, Germany, employing the narrowband 3GPP Adaptive Multirate (AMR-NB) speech codec (3GPP, 1999).

For each recognition experiment, an HMM-based phoneme recognition engine was trained using the Cambridge Hidden Markov Model Toolkit (HTK). In all cases, the HMM set consisted of 48 acoustic models according to the simplified TIMIT phone set described in (Lee and Hon, 1989). Each HMM consisted of three states with a left-to-right topology and 16 Gaussian densities per state. For feature extraction, the standard HTK frontend was used, producing 39-element feature vectors at a rate of $100\ \mathrm{s}^{-1}$. The features used were mel-frequency cepstral coefficients (MFCCs) including C0 and first- and second-order derivatives. The time-domain pre-processing of the frontend comprised framewise mean subtraction, pre-emphasis ($\alpha = 0.97$), and the application of a 25 ms Hamming window. The mel-filterbank consisted of 26 triangular bandpass filters equally distributed along the mel-frequency range corresponding to 0–8 kHz in the wideband speech case; in the

Exp.	ASR Frontend	Speech data	% ACC	% CORR
A1	wideband	TIMIT	66.72	74.95
A2		TIMIT–AMRWB1265(sim)	65.17	73.74
A3		WTIMIT = TIMIT–AMRWB1265(3G)	60.43	70.33
A4	narrowband	TIMIT↓2	65.27	74.00
A5		TIMIT↓2–AMRNB122(sim)	62.99	72.24
A6		TIMIT↓2–AMRNB122(3G)	59.22	69.46

Table 1: Results of phoneme recognition experiments with matched training and testing conditions.

Exp.	ASR Frontend	Speech data	% ACC	% CORR
B1	wideband	WTIMIT (HMM of exp. A2)	58.61	68.28
B2	narrowband	WTIMIT↓2 (HMM of exp. A6)	57.50	66.21

Table 2: Results of phoneme recognition experiments with mismatched training and testing conditions.

narrowband case, 23 filters along the range corresponding to 0–4 kHz were used. Cepstral mean normalization was applied to each file individually. All 48 HMMs were initialized identically using the global means and variances of the training partition of the respective corpus; thus, there was no dependence of the models on the time alignments provided with the TIMIT labels.

Phoneme recognition was performed on the test partitions of each corpus using HTK. A simple unrestricted phoneloop grammar – i. e., no language model – was employed. For the computation of the recognition rates, consecutive silences were counted as a single period of silence. Moreover, some equivalence classes of phones were defined according to (Lee and Hon, 1989), resulting in 39 phonemelike classes.

4.2. Experimental Results

Tables 1 and 2 show the results of several phoneme recognition experiments in matched and mismatched conditions, respectively. Recognition rates are given in percent accuracy (ACC) – i. e., considering deletions, substitutions, and insertions – and percent correctness (CORR) – i. e., ignoring insertions. Depending on the sampling rate of the respective speech data to be used for training and testing, the wideband or narrowband ASR frontend was employed.

4.2.1. Effects of Bandwidth, Codec, and Channel

Note at first the bandwidth effect of exp. A1 vs. A4 in Table 1. We observe that the wideband ASR frontend yields a $+1.45$ % higher phoneme recognition accuracy than the narrowband frontend. This represents a $+4.18$ % relative improvement of the phoneme error rate (PER). Similar observations are made for exp. A2 vs. A5 with a $+2.18$ % higher accuracy ($+5.89$ % rel. PER) and for exp. A3 vs. A6 with a $+1.21$ % higher accuracy ($+2.97$% rel. PER), respectively. Comparable bandwidth effects are also observed regarding the results in percent correctness.

Note further the codec and channel effects of experiments A1 vs. A2 and A1 vs. A3 in Table 1 using the wideband ASR frontend. While the pure codec influence in exp. A2 causes a -1.55 % change in accuracy (-4.66 % rel. PER), the influence of both codec and channel in exp. A3 alters the accuracy by -6.29 % (-18.90 % rel. PER). Similar observations are reported for experiments A4 vs. A5 and A4 vs. A6 using the narrowband ASR frontend. Here the changes in accuracy are -2.28 % due to the codec influence (-6.56 % rel. PER) and -6.05 % due to the codec and channel influence (-17.42 % rel. PER), respectively. Again, comparable effects are observed based on percent correctness. In consequence, we can state that just simulating the codecs gives a much too optimistic picture of real-world channels, both wideband and narrowband.

4.2.2. Practical Options

In practice, narrowband IVR systems are conceptually designed according to exp. A6, i. e., having narrowband telephony speech data for training and testing (matched conditions). But what about future wideband IVR systems? There are some practical options in order to deal with the incoming wideband speech. Table 3 summarizes the relative PER changes of these options related to the narrowband telephony baseline of exp. A6.

First of all, the 16 kHz input speech could be decimated again to 8 kHz sampling rate and recognized using existing narrowband acoustic models. However, this produces a mismatch between training and testing conditions, which is demonstrated by exp. B2 (Table 2). A performance degradation regarding exp. A6 vs. B2 is the consequence, showing in the -1.72 % change in accuracy (-4.22 % rel. PER, Table 3). Consequently, it would be even better to have a narrowband phone call according to exp. A6.

As a further practical solution, the 16 kHz input speech could be recognized using newly trained wideband acoustic models based on purely simulated AMR-WB speech data. Hence, channel effects are not considered in HMM training. This option is attractive since it does not require the transmission of training data over a real wideband telephone network. However, this again leads to a mismatch demonstrated by exp. B1 (Table 2). Indeed, the phoneme recognition accuracy of exp. B1 is $+1.11$ % higher than of exp. B2. However, it is still -0.61 % off the narrowband telephony baseline of exp. A6 (-1.50 % rel. PER, Table 3). Superior performance is achieved by exp. A3, which represents a true wideband IVR system employing transcoderfree operation (TrFO), as possible, e. g., in voice over IP (VoIP). However, it requires real wideband telephony speech data for HMM training in order to match the testing conditions appropriately. Based on our WTIMIT corpus in

Exp.	% rel. PER
A6	± 0.00
B2	-4.22
B1	-1.50
A3	$+2.97$

Table 3: Relative change of the phoneme error rate (PER) for experiments B2, B1, and A3 in relation to A6.

exp. A3, the phoneme recognition is thus $+1.21$ % more accurate in comparison to the narrowband telephony baseline, which represents a relative PER improvement of $+2.97$ % (Table 3). Again, the results in percent correctness follow comparable trends.

5. Conclusions

This paper presents a new corpus called WTIMIT. It has been derived from TIMIT by transmission over a 3G AMR-WB mobile network. The original TIMIT labels can largely be used for WTIMIT as well. Containing the effects of a real wideband telephone codec and channel, the WTIMIT corpus is suitable for development and evaluation of future wideband IVR systems. Phoneme recognition experiments show that matched training and testing conditions based on WTIMIT speech data improve the narrowband telephony baseline accuracy. Other practical options in order to deal with wideband telephony speech suffer from mismatched conditions and thus degrade the performance.

6. Database Distribution

The WTIMIT 1.0 corpus containing all transmitted TIMIT speech files, calibration tones, and documentation has been released by the Linguistic Data Consortium (LDC). It is being distributed via catalogue number LDC2010S02 (Bauer and Fingscheidt, 2009b). Additionally, it is planned to provide later releases with the complementary TIMIT speech data that has been transmitted over T-Mobile's 3G mobile telephony network in Braunschweig, Germany, employing the AMR-NB speech codec (see Section 4).

If you have any questions, problems, or suggestions concerning WTIMIT, please let us know by sending an email with your concerns to Tim Fingscheidt: fingscheidt@ifn.ing.tu-bs.de. We will be glad to assist you.

7. Acknowledgement

We would like to thank Mr. Dirk Kistowski-Cames, Deutsche Telekom AG, Bonn, Germany, for providing general project support and SIM cards, and Mr. Petri Lang, T-Mobile NL, The Hague, The Netherlands, for local support and even more SIM cards. Many thanks to Mr. Panu Nevala, Nokia, Oulu, Finland, for providing us with the prepared mobile phones, which are in that form not available on the market.

The work on our side was funded by the German Research Foundation (DFG) under grant no. FI 1494/2-1.

8. References

3GPP. 1999. Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions (3G TS 26.090).

3GPP. 2001. Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions (3GPP TS 26.190).

P. Bauer and T. Fingscheidt. 2009a. A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription. In *Proc. of EUSIPCO*, pages 1839–1843.

P. Bauer and T. Fingscheidt. 2009b. WTIMIT 1.0. Linguistic Data Consortium, Philadelphia, USA.

K. L. Brown and E. B. George. 1995. CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition. In *Proc. of ICASSP*, pages 105–108.

J. S. Garofolo et al. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia, USA.

J. S. Garofolo et al. 1996. FFMTIMIT. Linguistic Data Consortium, Philadelphia, USA.

ITU. 1988. ITU-T Recommendation G.722, 7 kHz Audio-Coding Within 64 kbit/s.

ITU. 1999. ITU-T Recommendation G.722.1, Coding at 24 and 32 kbit/s for Hands-free Operation in Systems With Low Frame Loss.

ITU. 2002. ITU-T Recommendation G.722.2, Wideband Coding of Speech at Around 16 kbits/s Using Adaptive Multi-Rate Wideband (AMR-WB).

ITU. 2008. ITU-T Recommendation G.711.1, Wideband Embedded Extension for G.711 Pulse Code Modulation.

C. Jankowski et al. 1990. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. In *Proc. of ICASSP*, pages 109–112.

K.-F. Lee and H.-W. Hon. 1989. Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648.

N. Morales et al. 2008. STC-TIMIT: Generation of a Single-channel Telephone Corpus. In *Proc. of LREC*, pages 391–395.

D. A. Reynolds. 1997. HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects. In *Proc. of ICASSP*, volume 2, pages 1535–1538.