

Improving ASR performance on PDA by contamination of training data

Christophe Ris and Laurent Couvreur

Multitel & FPMS-TCTS, Avenue Copernic 1, B-7000 Mons, Belgium

ris,couvreur@multitel.be

Abstract

Automatic Speech Recognition (ASR) on Personal Digital Assistant (PDA) suffers from the intrinsic hardware characteristics of the audio interface, for example, low quality microphones and device internal noises. In this paper, we propose to compensate for these weaknesses by contaminating clean training data with the distortion sources that are specific to the target device. We present a method to estimate both the frequency response of the audio acquisition channel and the internal additive noise from a few tens of minutes of recordings on PDA. The channel characteristics are estimated from the long-term power spectra of clean speech and PDA recordings, while the noise power spectrum is estimated during silence segments in these recordings. All the recordings are performed in a controlled way, *i.e.* quiet environment and no reverberation, in order to ensure that we measure only the internal device characteristics. The PDA-specific training data are then obtained by filtering the clean training data with the audio channel frequency response and contaminating them with internal noise, and a specific acoustic model is eventually trained for the target device. Recognition tests have been performed on digit sequences on three different PDA's. Our approach has been compared to other channel and noise robust methods and presents very competitive performance.

1. Introduction

The last few years have seen the huge development of ubiquitous devices (mobile phones, PDA, laptop computers, tablet computers, etc) and dedicated services (information, games, remote support, etc). Together with the commercial success of these devices, the connectivity and communication possibilities have also constantly increased in terms of performance and availability, allowing the potential applications to be more and more complex. As a consequence, the interaction between the humans and these applications has become a crucial re-

search domain and aims at optimally combining different interface modes such as keyboards, haptics, pens, voice, etc, according to the intrinsic capabilities of the mobile devices as small display, no keyboard, small computational capabilities, etc.

In such a framework, Automatic Speech Recognition (ASR) has become a major component of nowadays Human-Computer Interface (HCI), appearing as a natural way to interface with computers, improving the ergonomics of man-machine dialogues. However, the integration of accurate ASR is still a difficult problem as many sources of degradation can alter the speech signal and severely degrade the ASR performance. One of the source of degradation comes from the mobile equipments themselves that are generally equipped with low-quality audio hardware (microphones and analog-to-digital converter) whose design rarely takes into account automatic speech recognition. There exist various approaches to recover the performance, at least partly, for example channel compensation [2, 3, 4], noise reduction [5, 6, 7] or model adaptation [8, 9, 10]. Besides, it appears that ASR on degraded speech can reach quasi-optimal performance as compared to ASR on clean speech when the acoustic model is trained on data recorded in conditions similar to the operating conditions. Unfortunately, this implies to record a large amount of speech data directly on the target device which is generally not practical or even possible.

In this paper, we propose to simulate the last approach for ASR on PDA by contaminating clean training data with the sources of distortion specific to the target device, that is the audio acquisition channel filter and the internal additive noise. We present a method to estimate both the frequency response of the audio acquisition channel and the additive noise from a few tens of minutes of recordings on PDA. The paper is organized as follows. In section 2, we have a closer analysis of the degradation sources for speech recorded on PDA. In section 3, we describe our approach for estimating the channel filter and the internal noise on PDA, and contaminating the train-

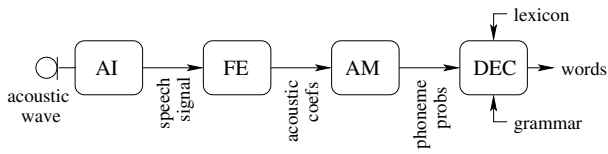


Figure 1: A typical ASR system: microphone, audio interface (AI), front-end (FE), acoustic model (MA) and word decoder (DEC).

ing data. Section 4 will present the results of ASR experiments on speech data recorded on PDA. Conclusions are drawn in section 5.

2. Problem Statement

A typical ASR system, as it is considered in this work, is depicted in figure 1. It consists of four main blocks. First, the audio interface converts the acoustic wave that is measured by a microphone into a digital speech signal. Second, the front-end (FE) chops the speech signal into frames and computes for each frame a set of acoustic coefficients that capture the essential shape of the power spectrum. In this work, the acoustic coefficient are obtained via the Perceptual Linear Predictive (PLP) algorithm [1]. Next, the acoustic coefficient vectors are fed into the acoustic model (MA) which estimates a probability score for every phoneme of the language under consideration. Here, the acoustic model is based on the Multi Layer Perceptron (MLP) / Hidden Markov Models (HMM) paradigm [11]. Such an acoustic model has to be trained *a priori* on large speech database containing a few hours of material. Finally, the word decoder (DEC) searches for the most likely word sequence, under the constraint of a phonetic lexicon and a word grammar, given the sequence of probability vectors for all the frames.

In our research, we are interested in testing such an ASR system on pocket computers. Actually, three PDA's are considered in this work (see figure 2). In order to avoid any direct comparison between these products, they will not be mentioned explicitly in the following. Instead, each of them will be associated with a dummy name of the form "PDA X" without defining to which device each such name actually corresponds. In all the cases, we have observed that the recognition performance degrades severely in comparison to the performance of the same system tested on a workstation for the same recognition tasks. It remains true even in laboratory con-



Figure 2: View of pocket computers: (a) Dell Axim X5[®], (b) HP Ipaq 5450[®] and (c) Symbol PDT 8100[®].

ditions, *i.e.*, noise-free and reverberation-free environments.

In order to explain this observation, we derive the following mathematical framework. Define x_n as the discrete-time speech signal that is delivered by the PDA audio interface to the front-end block of the ASR system. As we stated earlier, the front-end block will process x_n in order to extract the time evolution of its power spectrum. To do so, the very first step consists in computing its Short Term Fourier Transform (STFT),

$$X_{m,k} = \sum_{n=-\infty}^{\infty} w_{n-m} x_n z^{-nk} \quad (1)$$

with $z = e^{-j2\pi/N}$. Every coefficient $X_{m,k}$ is intended to estimate the spectrum of the speech signal at the m -th time location for the k -th discrete frequency $\omega_k = 2\pi F_r k/N$ with F_r being the sampling rate, 8 kHz in this work. It is obtained by first applying a window function w_n to the speech signal and next computing the Discrete Fourier Transform (DFT) of the windowed signal. The window function has a finite support of length N , *i.e.* $w_n = 0$ for $n < 0$ and $n > N - 1$, vanishing smoothly at its ends. In this work, a Hanning window is used. Its length is set equal to 240 samples, *i.e.* 30 ms at 8 kHz, as a tradeoff between ensuring the stationarity of the speech signal within the window and providing a high enough frequency resolution. The STFT coefficients are classically computed at regular times intervals. Here, they are obtained every 80 samples, *i.e.* 10 ms at 8 kHz. The power spectrum $|X_{m,k}|^2$ is eventually obtained by taking the square of the magnitude of the spectrum coefficients.

If we assume that the audio interface behaves like a linear time-invariant system, it is entirely characterized by its impulse response h_n . If we further assume that it

generates some internal noise v_n , we can write

$$x_n = h_n * s_n + v_n = \sum_{l=-\infty}^{\infty} h_{n-l} s_l + v_n \quad (2)$$

where s_n denotes an hypothetical speech signal as it would be measured by an ideal audio interface in a noise-free and reverberation-free environment. By taking the STFT of both sides, we obtain

$$\begin{aligned} X_{m,k} &= \sum_{n=-\infty}^{\infty} w_{n-m} \left(\sum_{l=-\infty}^{\infty} h_{n-l} s_l + v_n \right) z^{-nk} \\ &= \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} w_{n-m} h_{n-l} s_l + V_{m,k} \end{aligned} \quad (3)$$

where $V_{m,k}$ stands for the spectral coefficients of the internal noise signal. By making the change of variables $n' = n - l$ and interchanging the summation order, we can further develop equation (3),

$$X_{m,k} = \sum_{l=-\infty}^{\infty} s_l z^{-lk} \sum_{n'=-\infty}^{\infty} w_{n'+l-m} h_{n'} z^{-ln'} + V_{m,k}. \quad (4)$$

If we assume that the impulse response of the audio interface is causal and short compare to the length N of the window function such that w_n is approximatively constant over the duration of h_n , then we arrive at the following equation,

$$\begin{aligned} X_{m,k} &\simeq \sum_{l=-\infty}^{\infty} s_l z^{-lk} \sum_{n'=-\infty}^{\infty} w_{l-m} h_{n'} z^{-n'k} + V_{m,k} \\ &= \sum_{l=-\infty}^{\infty} w_{l-m} s_l z^{-lk} \sum_{n'=-\infty}^{\infty} h_{n'} z^{-n'k} + V_{m,k} \\ &= S_{m,k} H_k + V_{m,k} \end{aligned} \quad (5)$$

where $S_{m,k}$ stands for the spectral coefficients of the hypothetical speech signal s_n and H_k is the frequency response of the audio interface.

Since we are interested in the power spectrum, we take the square of both sides of equation (5),

$$\begin{aligned} |X_{m,k}|^2 &= |S_{m,k}|^2 |H_k|^2 + |V_{m,k}|^2 \\ &\quad + S_{m,k} H_k V_{m,k}^* + S_{m,k}^* H_k^* V_{m,k}. \end{aligned} \quad (6)$$

In practice, the speech signal and the internal noise are considered to be statistically independent. Hence, the last two terms are classically assumed to be null though this assumption is true only the mean sense. We finally model the impact of the audio interface on the speech signal by the following equation

$$|X_{m,k}|^2 = |S_{m,k}|^2 |H_k|^2 + |V_{m,k}|^2. \quad (7)$$

This equation is central to our problem and reads that the power spectrum $|X_{m,k}|^2$ of the speech signal results from two components, first the power spectrum $|S_{m,k}|^2$ of the speech source altered by the audio channel $|H_k|^2$, and secondly the power spectrum $|V_{m,k}|^2$ of the internal noise. Clearly, two distinct audio interfaces are likely to have different characteristics, hence distorting the speech source in different ways.

It is common to visualize the time evolution of the power spectrum as a spectrogram, which consists in a three-dimensional representation with the time as abscissa, the frequency as ordinate and the power intensity as a colormap. Figure 3.(a) shows the spectrogram of the utterance ‘‘zéro deux sept’’ (‘‘027’’ in French) recorded on a workstation equipped with a studio-grade microphone and a high-quality sound board. Figure 3.(b) shows the spectrogram of the same utterance recorded on PDA 3. Though both utterances were recorded simultaneously, we clearly observe significant differences between their spectrograms. The reasons for these discrepancies are unclear. They may result from low-quality electronics, too severe anti-aliasing filter or acoustical interferences at sound holes in the pocket computer external case. Nevertheless, they are responsible for the degradation of ASR performance on PDA because the acoustic model is classically trained on speech material recorded with a high-quality audio interface. During training, it learns how to map some spectral characteristics to some phonemes. When used on PDA, the same phonemes will correspond to different spectral characteristics, or the same spectral characteristics will correspond to other phonemes. Consequently, the acoustic model produces unreliable probability vectors and the decoding search is misled to incorrect recognition results on PDA.

3. Proposed Method

Many approaches have been developed in order to reduce the mismatch between the spectral characteristics of the training speech and the ones during operation. They can generally be cast into two categories, namely compensation methods and adaptation methods. In the former case, the corrupted speech signal or any of its representation in the ASR process before the acoustic model block is compensated for the effect of the audio interface channel and the internal noise such that the source speech signal is restored, keeping the acoustic model as it is. In the latter case, the corrupted speech signal is not modified but the acoustic model is adapted to it. Well-

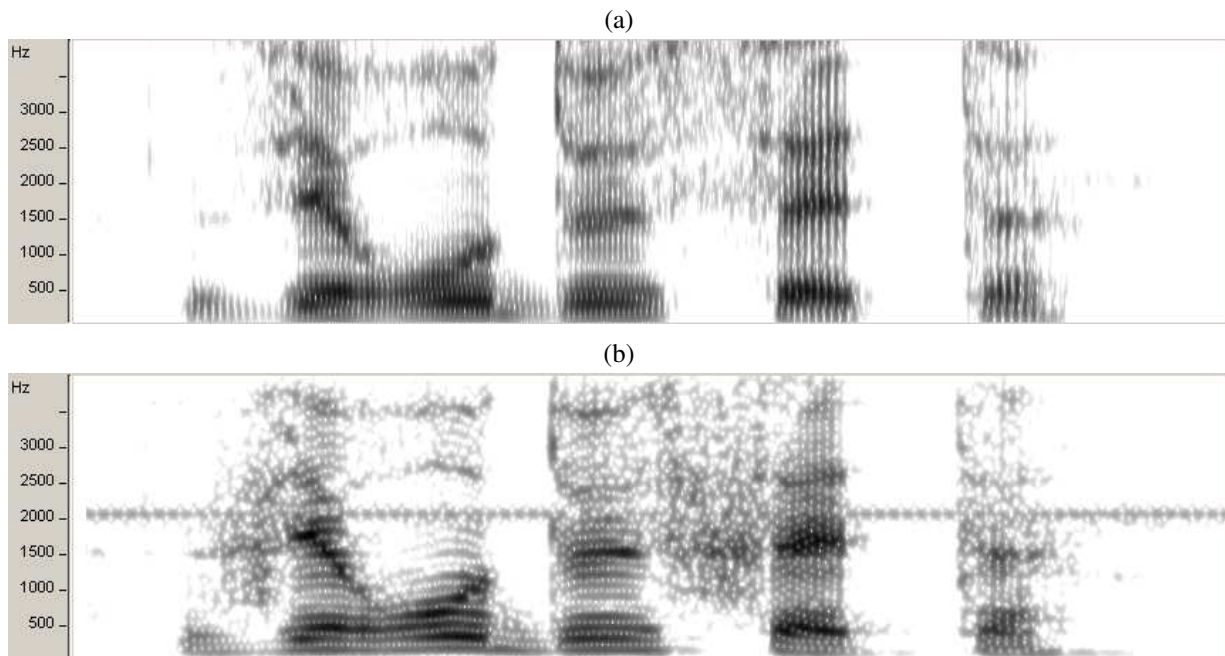


Figure 3: Spectrogram for the utterance “zéro deux sept” (“0 2 7” in French) recorded simultaneously on (a) a workstation with a studio-grade microphone and a high-quality sound board, and (b) a pocket computer PDA 3.

known techniques for channel compensation are Relative SpecTrAl (RASTA) filtering [2, 4] and Cepstral Mean Subtraction (CMS) [3, 4]. These techniques consists in applying a non-linear transformation to the power spectrum such that multiplication in equation (7) becomes addition and operands can be separated. Noise compensation methods typically rely on the estimation of the noise power spectrum during non-speech segments and subtraction from the corrupted power spectrum [5, 7]. Classical adaptation techniques are Maximum Likelihood Linear Regression [9], Parallel Model Compensation [8], ... Note that these methods are hard to work out for hybrid MLP/HMM ASR systems.

In this paper, we suggest to specialize the acoustic model to the characteristics of the PDA audio interface in order to improve the ASR performance. By specialization, we mean training the acoustic model on data recorded in conditions similar to the operating conditions. Our approach can be viewed as a kind of adaptation method except that the acoustic model is not just slightly modified but reestimated from scratch. Since it is not convenient to record a specific training speech database on every PDA, we suggest that it can be obtained by contaminating an existing training speech database, which was collected in noiseless and anechoic conditions via a high-quality audio interface, with the audio interface characteristics of the PDA under consideration. To do so,

the frequency response of the audio interface as well as the internal noise have to be estimated.

Direct measure of the frequency response requires a specific equipment and a rigorous protocol. For practical reasons, it could be easier to estimate it from speech recordings. As we explained earlier, the audio interface acts as a filter attenuating some parts of the speech power spectrum and enhancing other ones. We claim that the information about the frequency response of the audio interface that is relevant for the ASR process can be extracted from the Long-Term Spectrum (LTS) of speech recordings. Given a speech signal x_n , its LTS coefficient \bar{X}_k for the k -th discrete frequency is defined as the power spectrum $|X_{m,k}|^2$ averaged over time, that is,

$$\bar{X}_k = \frac{1}{N_x} \sum_{m=0}^{N_x-1} |X_{m,k}|^2 \quad (8)$$

with N_x denoting the number of analysis frames. Note that a speech activity detector is used in order to cancel out silence frames and estimate the LTS from frames where speech dominates the internal noise. Based on the assumption that the performance of ASR systems are all the better as the LTS of data used for the training of the acoustic model are similar to the LTS of data encountered during the recognition task, we propose a method to prepare the training speech data by adequately modifying their LTS.

First, one chooses a speech database for training purpose and records speech data with the PDA to be used. The training material is typically obtained with a high-quality audio interface while the PDA material may be corrupted by some severe distortions as we explained earlier. Note that the PDA recordings are performed in a quiet non-reverberant environment such that only the characteristics of the device acquisition hardware affect the signal. Then, the LTS \bar{X}_k^{Train} of speech data dedicated to training the acoustic model is computed,

$$\bar{X}_k^{Train} = \frac{1}{N_x^{Train}} \sum_{m=0}^{N_x^{Train}-1} |X_{m,k}^{Train}|^2. \quad (9)$$

Likewise, the LTS \bar{X}_k^{PDA} is computed from some speech material that is recorded with the PDA under consideration,

$$\bar{X}_k^{PDA} = \frac{1}{N_x^{PDA}} \sum_{m=0}^{N_x^{PDA}-1} |X_{m,k}^{PDA}|^2. \quad (10)$$

One question of interest is what the recordings should contain in order to provide a reliable LTS estimate. We can say that there should be a sufficient number of speakers and the vocabulary should be large enough such that the speech data will cover satisfyingly the acoustic variabilities. Another question of interest is how long the recordings should be if the speaker and vocabulary conditions are satisfied. It is known that the mean estimator of equation (8) is consistent, *i.e.* the more data the better the estimate, yet there is a critical amount of data that is required to ensure a reliable LTS estimate. Figure 4 shows the evolution of the normalized mean square error (in percent) between two successive estimations of the LTS for a recording obtained on PDA 2. These estimations are produced at 1 minute intervals from a 180 minute recording. As we can see, the error decreases as more data are used to compute the LTS estimate, falling below 1% and stabilizing after 30 minutes of recording. Note that the duration is given for the complete recorded signal, *i.e.* including the silence frames that represent about 23% of all the frames for our recordings.

Secondly, a mapping function $\mathcal{F}_k^{Train/PDA}$ is derived from the LTS estimates \bar{X}_k^{Train} and \bar{X}_k^{PDA} ,

$$\mathcal{F}_k^{Train/PDA} = \frac{\bar{X}_k^{PDA}/E_X^{PDA}}{\bar{X}_k^{Train}/E_X^{Train}} \quad (11)$$

where E_X^{Train} and E_X^{PDA} stand for the long-term average of the frame energy of the signal recorded with the high-quality audio interface and the PDA, respectively. The mapping function is next smoothed by applying a

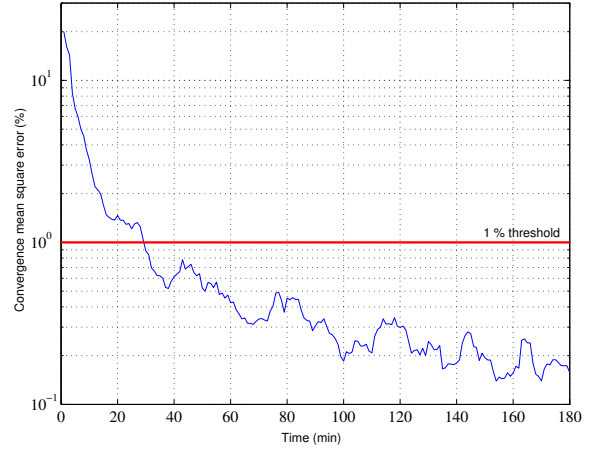


Figure 4: Evolution of the normalized mean square error (in percent) between two successive estimations of the LTS for a recording obtained on PDA 2.

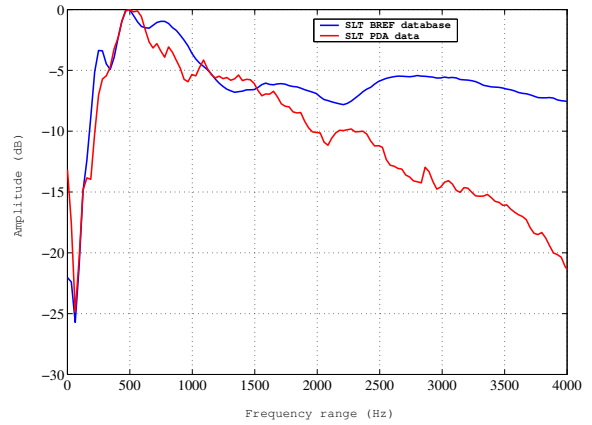


Figure 5: Comparison between LTS of two sets of speech data: high-quality audio interface vs. PDA 1.

mean filter of third order in the log domain,

$$\mathcal{F}_k^{Train/PDA} = \exp \left(\left[\log \mathcal{F}_{k-1}^{Train/PDA} + \log \mathcal{F}_{k+1}^{Train/PDA} \right] / 3 \right) \quad (12)$$

As an example, figure 5 displays the long term spectra of 30 minutes of read speech in French recorded with a high-quality microphone and downsampled at 8 kHz, and the corresponding speech data recorded on PDA 1. The speech LTS is naturally low-pass with a bulk of energy below 1 kHz and decreasing gently for higher frequencies. We observe that the speech LTS for the PDA is severely attenuated over 2 kHz denoting the strong low-pass effect of its audio interface. Figure 6 shows the mapping function that is derived from the two LTS of figure 5 using equation (11).

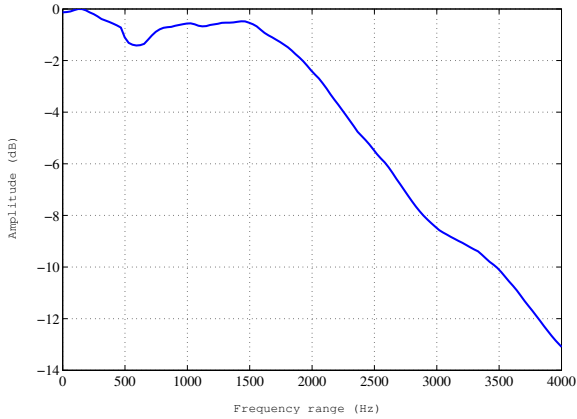


Figure 6: *Function mapping LTS of a high-quality microphone database towards LTS of speech material recorded on PDA 1.*

Finally, the mapping function is used to contaminate the training speech data. This is done by inserting the mapping function in the front-end when computing the acoustic vectors for the training database: element-wise multiplication is performed between the power spectrum of every analysis frame and the mapping vector,

$$|X_{m,k}^{Train/PDA}|^2 = |X_{m,k}^{Train}|^2 \mathcal{F}_k^{Train/PDA}, \forall k. \quad (13)$$

In our example (see figure 5), it has the effect of attenuating significantly the power spectrum over 2 kHz, hence better reflecting the power spectrum as it would have been observed on the PDA. Once the whole training database has been processed, an acoustic model that is more representative of the PDA audio interface can be trained as usually done.

The data contamination approach can be also used for compensating the internal additive noise. Indeed, under the hypothesis that this noise is stationary, its spectral characteristics can be extracted during the silence sections in the PDA recordings, that is, the frames that were cancelled out by the speech detector when estimating the mapping function. This estimated noise signal is therefore added to the clean speech training set. The acoustic model trained with these contaminated data will be inherently more robust to the specific device noise.

Our approach is possible because we assume that the characteristics of the PDA audio interface are time-invariant and can be modeled once for all. In any case, our approach is robust to more difficult noise and acoustic distortion like environmental noise or room reverberation. Environmental noise is typically time-varying and it would be hard to capture representative data for contam-

ination. Besides, impulse responses corresponding to reverberation are highly varying and always longer than the analysis frame length, which makes the model of equation (7) fail.

4. Experimental Results

4.1. Speech Database

In order to assess the approach described in the previous section, we have performed ASR tests on sequences of digits in French.

The speech material for training the acoustic model comes from the BDSOONS database [14], which consists of connected digit sequences in French among others. The speech signals from this database were downsampled at 8 kHz.

The test set was recorded simultaneously on three PDA's (see figure 2) and a workstation equipped with high-quality audio interface. It contains 1000 utterances that consist in sequences of 3 to 6 digits in French. They were recorded by 3 speakers in a noise-free and low-reverberation enclosure such that no other effect than the internal characteristics on the audio interfaces affects the speech signals. The PDA's and the high-quality microphone were all located within arm's reach in front of the speaker. All the recordings were performed at 8 kHz.

We chose a subset of the BREF [13] database, a large vocabulary corpus of read speech in French with a high speaker diversity and phonetic coverage, for estimating LTS and deriving the mapping functions. To do so, for every PDA, utterances were selected randomly, played back with a studio-grade loudspeaker in a noise-free and reverberation-free environment and simultaneously recorded with the PDA's. All recording were performed at 8 kHz.

4.2. Audio channel compensation

First, we would like to verify that the distortion model of equation (3) is valid. More especially, we want to check whether the PDA impulse responses are short enough with respect to the length of the analysis frame in the front-end block of the ASR process. By its very definition, an impulse response can be measured by producing an impulsive sound and recording the response signal at the PDA. In practice, it is hard to deliver a high (ideally infinite) energy in a very (ideally infinitely) short time. Gun shot or ballon blowup are sometimes used, we preferred the Time-Stretched Impulse TSP method [12]. It consists in driving a loudspeaker with a chirp signal that

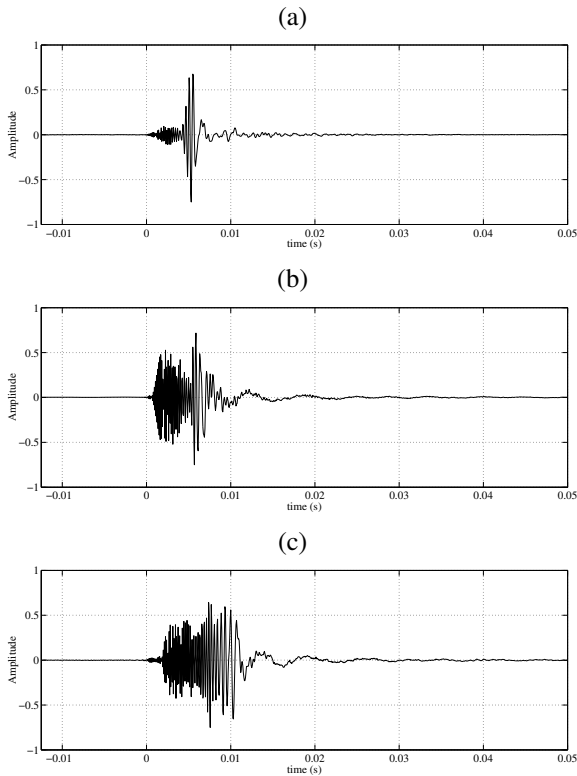


Figure 7: Impulse response of (a) PDA 1, (b) PDA 2 and (c) PDA 3 audio interfaces as measured via the Time-Stretched Pulse method.

spreads its energy from high frequencies to low frequencies linearly over time. The TSP response is simultaneously recorded on the PDA and then convolved with the inverse TSP to derive the impulse response. Figure 7 shows the impulse for the three PDA’s located at half a meter from the loudspeaker in an anechoic room. Clearly, they are shorter than the length of the analysis frame, namely 30 ms. Hence, we can consider that the model of equation (3) is valid for the PDA audio interfaces under the assumption that they behave as linear time-invariant systems.

One can suggest that if we are able to measure the impulse response of a PDA audio interface, it should be used to contaminate the training database. In our opinion, the measure of an impulse response is by far more laborious than simply recording speech signals on the PDA. Hence, we believe that our speech-based approach for estimating the PDA frequency response is more natural and simpler to implement reliably.

We have compared the approach by contamination of the training data with two standard procedures for channel compensation, namely RASTA filtering and CMS compensation. In all the cases, the basic acoustic features

Table 1: ASR word error rates for the “mapping-based” channel compensation technique and comparison with two standard channel compensation techniques: RASTA filtering and CMS.

PDA model	PLP	RASTA-PLP	CMS-PLP	LTS map.
PDA 1	7.7%	5.3%	3.9%	3.5%
PDA 2	4.0%	3.7%	2.6%	2.4%
PDA 3	24.3%	7.5%	6.3%	7.8%

were PLP coefficients. Note that in the case of the LTS mapping, only the training data are modified and standard PLP coefficients without any mapping are used for the test data. Table 1 presents the results that we obtained in terms of word error rates. Note, as a reference, that the baseline ASR performance for connected digits recorded in a quiet environment with a high-quality microphone is 0.8% word error rate.

First, we observe that the degradation of the recognition performance compared to high-quality recordings are very dependent on the type of PDA. The very poor performance of the PLP coefficients on PDA 3 can be partially explained by the presence of an internal additive noise at 2 kHz. This problem will be addressed in next section. We note also the better performance of cepstral mean subtraction compared to the RASTA filtering. Performance of the LTS mapping are very competitive with the standard channel compensation approaches. Note that the contamination approach and the channel compensation methods are conceptually opposite and, therefore, cannot be combined.

4.3. Internal noise compensation

As mentioned above, we have observed that the signal recorded with PDA 3 is corrupted with a narrow band noise at 2 kHz. We have estimated the spectral characteristics of this noise and artificially corrupted the training speech data with this noise. This approach is compared to a classical noise reduction technique, namely Wiener filtering [5]. Table 2 presents results of the different combination of channel robust techniques, namely, RASTA filtering, CMS and LTS mapping, and additive noise robust techniques, namely Wiener filtering and data contamination for PDA 3.

We see that, as for the effect of the channel, the noise contamination of the training data gives very competitive results compared to a classical denoising technique. Here

Table 2: ASR word error rates for combinations of noise robust and channel robust methods. Comparison between compensation and contamination approaches. Results for the PDA 3.

Methods	None	Wiener filt.	Noise contam.
None	24.3%	3.6%	3.9%
RASTA	7.5%	4.3%	3.8%
CMS	6.3%	2.1%	2.1%
LTS map.	7.8%	1.8%	2.2%

again, only the training data are modified, acoustic features for the test data are PLP. Note also, that this method makes the strong assumption that the spectral characteristics of the noise are time invariant which, in the case of a device internal noise, is a reasonable assumption.

5. Conclusions

In this paper, we have proposed an alternative approach to the specific problem of ASR on PDA, which consists in modifying the speech training data used to train the acoustic models, in such a way that they better fit the intrinsic characteristics of the PDA speech acquisition device. The idea consists in extracting the audio channel frequency response and the spectral content of the device internal noise from a few tens of minutes of speech recorded on the target PDA. The ASR experiments we carried out have shown very competitive results compared with classical channel compensation and noise subtraction methods. Note that it is not required to have the same recordings for both the extraction of the long-term spectrum (and therefore the mapping function) and for the training of the acoustic model. In our case, BREF was used for the mapping, while BDSOONS was used for training. Note also, that the acquisition procedure for the PDA is rather simple as a mere playback of 30 min of speech data in a controlled way (high-quality speakers, noise-free, reverberation-free environment) gave us very good results. Note finally that, although the approach is presented in the framework of a hybrid HMM/MLP system, it is not limited to that specific architecture.

6. References

[1] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *J. Acous. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[2] H. Hermansky and N. Morgan, "RASTA Processing of

Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[3] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] X. Huang, A. Acero and H.-W., Hon, "Spoken Language Processing: A guide to Theory, Algorithm, and System Development", *Prentice Hall*, pp. 522–525, 2001.

[5] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. of the IEEE*, vol. 67(12), pp. 1586–1604, 1979.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32(6), pp. 1109–1121, 1984.

[7] P. Lockwood and J. Boudy, "Experiments with a Non-Linear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", *Speech Communication*, vol. 22, pp. 1–15, 1992.

[8] M.J.F. Gales and S. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise.", *Proc. of ICASSP'92*, pp. 233–236, San Francisco (CA), 1992.

[9] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation", *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[10] J. Neto et al., "Speaker-Adaptation for Hybrid HMM/ANN Continuous Speech Recognition System", *Proc. of Eurospeech'95*, Madrid, 1995.

[11] H. Bourlard and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", *Kluwer Academic Publisher*, 1994.

[12] Y. Suzuki, F. Asano, H.-Y. Kim and T. Sone, "An optimum Computer-Generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses", *J. Acous. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, Feb. 1995.

[13] Lamel L.F., Gauvain J.L. and Eskénazi M., "*BREF, a Large Vocabulary Spoken Corpus for French*", EuroSpeech 1991, pp. 505-508, Geneva, Italy

[14] Carré R., Descout R., Eskénazi M., Mariani J. and Rossi M., "*The French Language Database: Defining, Planning and Recording a Large Database.*", ICASSP 1984, San Diego, California.