The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments

Ram H. Woo, Alex Park, and Timothy J. Hazen

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts, USA
emails: {*malex,hazen*}*@csail.mit.edu*

Abstract

In this paper we discuss data collection and preliminary experiments for a new speaker verification corpus collected on a small handheld device in multiple environments using multiple microphones. This corpus, which has been made publically available by MIT, is intended for explorations of the problem of robust speaker verification on handheld devices in noisy environments with limited training data. To provide a set of preliminary results, we examine text-dependent speaker verification under a variety of cross-conditional environment and microphone training constraints. Our preliminary results indicate that the presence of noise in the training data improves the robustness of our speaker verification models even when tested in mismatched environments.

1. Introduction

As technological improvements allow for the development of more powerful and ubiquitous handheld devices, such as PDAs and handheld computers, there exists a need for greater security as these devices may contain a myriad of sensitive or personal information. One potential option is the use of speaker verification technology using user specified pass-codes for secure user logins. Speaker verification provides a biometric layer of security that can enhance the security offered by personal identification numbers or user selected passwords. The focus of this work is to investigate issues of robustness for speaker verification using handheld devices.

Previous work on security-based speaker verification systems largely falls within two major domains: vestibule security and telephone-based user verification. Vestibule security focuses on the fast and secure physical access to restricted locations. Speaker verification allows for contact-less activation and mitigates the risks of stolen or lost keys, passwords, or keycards. Examples include work by Morin and Junqua [1] as well as Doddington [2]. Furthermore, speaker verification can be used in conjunction with other modalities (fingerprint, keypad, and/or face verification) to maximize flexibility, convenience, and performance [3, 4].

Telephone-based verification systems have a number of applications, particularly for conducting transactions requiring secure access to financial information (e.g., credit card information, bank account balance, etc) or other sensitive customer information (e.g., healthcare records). Examples include work by Bimbot *et al* [5] and Lamel, Barras, and Gauvain [6, 7].

One of the challenges in implementing speaker verification on handheld devices arises from their greatest attribute: mobility. Unlike vestibule security systems, handheld devices are often used in highly variable acoustic environments such as quiet offices, busy cafeterias, or loud street intersections. In each environment, variations in the acoustic conditions and background noises will corrupt the speech signal leading to intra-speaker variability that can reduce the accuracy of speaker verification systems. The variability in microphones used with handheld devices can also have a substantial impact on performance in speaker verification systems.

Unlike test systems developed for use in feasibility studies, real world systems are further constrained by usability issues. One of the foremost concerns is ease of use. It is desirable for handheld based verification systems to allow for the quick and easy enrollment of new users, preferable within one short enrollment session. However, collecting only limited amounts of enrollment data can further damage the potential robustness of the system by limiting both the amount of available training material and its variability (both in environment and microphone).

In this paper we describe a pilot study that we conducted in the area of speaker verification on handheld devices in variable environments using variable microphones and limited amounts of enrollment data. In Section 2, we discuss the collection of a new corpus for studying this problem. We describe our speaker verification system in Section 3, present experiments and results in Section 4, and make concluding remarks in Section 5.

2. Data Collection

For our data collection, a prototype handheld device provided by Intel was used. In order to simulate scenarios encountered by real-world speech verification systems, the collected speech data consisted of two unique sets: a set of *enrolled* users and a different set of dedicated *imposters*. For the enrolled user set, speech data was collected over the course of two different twenty minute sessions (one for training and one for evaluation) that occurred on separate days. For the imposter set, users participated in a single twenty minute session.

In order to capture the expected variability of environmental and acoustic conditions inherent with the use of a hand-held device, both the environment and microphone conditions were varied during data collection. For each session, data was collected in three different locations (a quiet office, a mildly noisy lobby, and a busy street intersection) as well as with two different microphones (the built-in internal microphone of the handheld device and an external earpiece headset) leading to 6 distinct test conditions. By recording in actual noisy environments this corpus does contain the Lombard effect (i.e., speakers alter their style of speech in noisier conditions in an attempt to improve intelligibility). The Lombard effect is missing in corpora that simply add noise electronically to data collected in quiet environments.

Office/Earpiece	Lobby/Earpiece	Intersection/Earpiece
alex park	alex park	alex park
rocky road	chocolate fudge	mint chocolate chip
ken steele	ken steele	ken steele
rocky road	chocolate fudge	mint chocolate chip
thomas cronin	thomas cronin	thomas cronin
rocky road	chocolate fudge	mint chocolate chip
sai prasad	sai prasad	sai prasad
rocky road	chocolate fudge	mint chocolate chip
trenton young	trenton young	trenton young
Office/Internal	**Lobby/Internal**	**Intersection/Internal**
alex park	alex park	alex park
peppermint stick	pralines and cream	chunky monkey
ken steele	ken steele	ken steele
peppermint stick	pralines and cream	chunky monkey
thomas cronin	thomas cronin	thomas cronin
peppermint stick	pralines and cream	chunky monkey
sai prasad	sai prasad	sai prasad
peppermint stick	pralines and cream	chunky monkey
trenton young	trenton young	trenton young

Table 1: A sample list of phrases spoken in each session.

Example spectrograms of four different recording conditions are displayed in Figures 3 through 6 at the end of this paper. In examining the spectrograms one should notice the obvious low pass filtering characteristic of the ear-piece microphone (Figures 3 and 4) relative to the device's internal microphone (Figures 5 and 6). Also note the significant background noise present on the data collected on the noisy street corner (Figures 4 and 6). The spectrograms also show two intermittent noise artifacts caused by the device. A high frequency buzzing tone is evident in Figure 5, and a click sound at the onset of recording is evident in Figures 5 and 6.

Within each data collection session, the user recited a list of name and ice cream flavor phrases which were displayed on the hand-held device. An example phrase list is shown in Table 1. In total, 12 different list sets were created for enrolled users while 7 lists were created for imposters. Enrolled users recited two phrase lists which were almost identical, differing only in the location of the ice cream flavor phrases on the lists. The first phrase list was read in the enrolled user's initial data collection session, while the second list phrase was used in the subsequent follow-up session.

One criterion in the design of the data collection was to allow for a variety of cross-condition experiments across background environments and microphones. The ice cream flavor phrases were each read four times within a unique environment/microphone condition, thus allowing the investigation of text-dependent verification within various training/testing cross-condition cases. A second design criterion was the collection of phonetically rich data which could potentially be used within a speaker verification background model. The name phrases were selected to provide phonetic coverage. Each name phrase is read once in each environment/microphone condition thus allowing the option to also perform text-dependent speaker verification using multi-style trained models with these phrases.

In total, each session yielded 54 speech samples per user. This yielded 5,184 examples from enrolled users (2,592 per session) and 2,700 imposter examples from users not in the enrollment set. Within the enrolled set of 48 speakers, 22 were female while 26 were male. For the imposter set of 40 speakers, 17 were female while 23 were male.

3. Speaker Verification System

3.1. Verification Framework

Speaker verification systems for security purposes typically employ a text-dependent approach where enrolled users utter either a specific pass-phrase or a string of prompted digits. Word or phrase specific hidden Markov models (HMMs) for each enrolled user are typically used to perform the verification. In our experiments we use our own ASR-dependent speaker verification approach [8]. In this approach, a speech recognition engine is used to phonetically time-align an expected pass-phrase utterance. A phone-adaptive scoring mechanism is then employed to score the phonetic components of the utterance against speaker-dependent, phone-dependent models created during the enrollment phase. Details of this process can be found in [9]. Analytically, the speaker score, $Y(\mathbf{X}, S)$ for speaker S uttering the test phrase X can be described as:

$$Y(\mathbf{X}, S) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x} \in \mathbf{X}} \log \left[\lambda_{S,\phi(\mathbf{x})} \frac{p(\mathbf{x}|S, \phi(\mathbf{x}))}{p(\mathbf{x}|\phi(\mathbf{x}))} + (1 - \lambda_{S,\phi(\mathbf{x})}) \frac{p(\mathbf{x}|S)}{p(\mathbf{x})} \right] \quad (1)$$

Here, $\mathbf{X}$ represents the set of feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, S is the purported speaker, and $\phi(\mathbf{x})$ is the phonetic model label for feature vector $\mathbf{x}$. The interpolation factor $\lambda_{S,\phi(\mathbf{x})}$ controls the relative weight that the scoring function assigns to phone dependent and phone independent models for each feature vector $\mathbf{x}$ for speaker S. Note that the speaker specific models are also normalized with speaker independent background models. The interpolation weight is determined as follows

$$\lambda_{S,\phi}(\mathbf{x}) = \frac{n_{S,\phi(\mathbf{x})}}{n_{S,\phi(\mathbf{x})} + \tau} \quad (2)$$

where $n_{S,\phi(\mathbf{x})}$ refers to the number of times the phonetic event $\phi(\mathbf{x})$ was observed in the enrollment data for speaker S, and τ is an empirically determined tuning parameter that was the same across all speakers and phones ($\tau = 5$ for our experiments). For scenarios involving limited enrollment data, this approach allows the system to smooth phone dependent models that may not be adequately trained with more robustly trained global Gaussian mixture models (GMMs) [10]. Experiments in [11] show that global GMMs outperform phone dependent GMMs in our limited enrollment data experiments, but our phone adaptive interpolation scheme offers modest improvements over global GMMs.

3.2. Acoustic Features

Because our experiments utilize the segment-based SUMMIT speech recognition system [12], we have used segment-based acoustic feature vectors for speaker verification as well. We have experimented with segment models using feature vectors comprised of Mel-frequency scale cepstral coefficient (MFCC) averages within fixed regions of hypothesized segments, with landmark models using features comprised of MFCC averages over regions surrounding hypothesized phonetic landmarks, and also with standard frame based models. In preliminary experiments in [11], a system based on a roughly equal weighting of segment and landmark models performed best, though further study of potential features to use in our framework would be worthwhile. In the experiments in this paper, our phone dependent approach uses context independent segment models and diphone landmark models derived from averages of 24-dimension MFCC frames.

4. Experimental Results

4.1. Evaluation Scenario

In our experiments, our system assumes that the user is speaking the enrollment pass-phrase of the purported user. Thus, the recognizer does not verify that the correct pass-phrase is spoken, but instead simply time-aligns the speech against the expected pass-phrase. We evaluate under the condition where the dedicated imposters speak the pass-phrases of the purported users they are attempting to impersonate. This yields a verification test for the scenario where both the user's device and pass-phrase have been stolen. In [11], it is shown that considerably better verification performance is achieved under the condition that an imposter does not know the user's pass-phrase and instead utters a random phrase that contains different phonetic content than the actual pass-phrase.

4.2. Preliminary Mismatched Condition Experiments

Our preliminary experiments with our new corpus explored the effects of mismatched training and testing conditions on system performance. In particular, we examined the impact of environment and microphone variability inherent with handheld devices. Figure 1 provides a preliminary glimpse of the impact of environment and microphone mismatches. For these experiments, users enrolled by repeating a single ice cream phrase four times in a particular environment/microphone condition. During testing, both enrolled users and dedicated imposters repeated the same ice cream flavor phrase. As can be seen, system performance varies widely as the environment or microphone is changed between the training and testing phase. While the fully matched trial (trained and tested in the office with an earpiece headset) produced an equal error rate (EER) of 9.4%, moving to a matched microphone/mismatched environment (trained in a lobby with the earpiece microphone but tested at a street intersection with an earpiece microphone) resulted in a relative degradation in EER of over 300% (EER of 29.2%). The mismatched microphone condition (trained in a lobby with an earpiece microphone and tested in a lobby with the device's internal microphones) also yielded a severe, though relatively smaller, degradation.

4.3. Cross-Environment Experiments

In order to examine the effects of environment variation independent of microphone variation we conducted a second set of experiments in which users were enrolled using five different name phrases spoken two times each (once with both the earpiece and internal microphones) during the initial enrollment phase.[1] The system was evaluated by testing the environment-specific models against test data collected in each of the three environments. In all tests, the phrases used in the enrollment session were identical to the phrases in the testing session. This task is fundamentally harder in comparison to the tests shown in Figure 1 as each name phrase is spoken only once for a given microphone/environment condition rather than 4 times. This is reflected in the higher EER of 13.75% obtained when training and testing in the office environment, as opposed to the EER of 9.38% experienced when we trained and tested solely on a single ice cream phrase uttered in the office/earpiece microphone condition. The results from these tests are compiled in Table 2. Full ROC curves for these experiments can be found in [11].

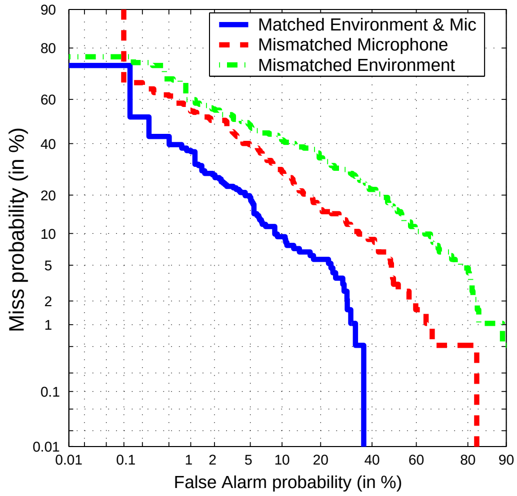

Figure 1: Detection-error tradeoff curves of preliminary cross-conditional tests.

Testing Location	Training Location: Office	Lobby	Intersection
Office	13.75%	13.33%	18.33%
Lobby	14.58%	14.79%	15.62%
Intersection	28.33%	30.00%	12.71%

Table 2: EERs of cross-conditional environment tests with models trained and tested in each of the three different environments.

Several interesting observations can be made from these results. In general, one would expect that the speaker verification system would have the lowest equal error rates in situations where the system is trained and tested in the same environmental conditions. However, when the speaker verification system was trained in the lobby environment, the system performed better when tested in the office environment (13.33%) than it did in the lobby environment (14.79%). When trained in the intersection environment, the speaker verification system proved most robust to variations in environment with a maximum cross-environment degradation of 5.65% (as compared cross-environment degradations of 14.58% and 16.67% for models trained in the office and lobby environments). Furthermore, despite the noisy conditions of the street intersection location, the train-intersection / test-intersection trial produced the lowest overall EER of 12.71%. Overall, it appears that the performance degradation experienced when moving from a "noisy" training environment to a "quiet" testing environment was not as severe as that of the reverse situation.

4.4. Cross-Microphone Experiments

Along with varied environments, speaker verification systems for mobile devices may be subject to variable microphone conditions (because interchangeable plug-in microphones are available for these devices). In order to understand the effect of mi-

[1] Variable names, rather than fixed ice cream flavor phrases, were used because each name phrase appeared in all of the six conditions while ice cream flavors each appeared in only one condition for a given phrase list. This limited the number of matched/mismatched environment and microphone tests that could be achieved with ice cream flavor phrases.

Testing Microphone	Training Microphone	
	Earpiece	Internal
Earpiece	11.11%	18.19%
Internal	22.36%	10.97%

Table 3: EERs of cross-conditional microphone tests with models trained and tested with each of the two microphones.

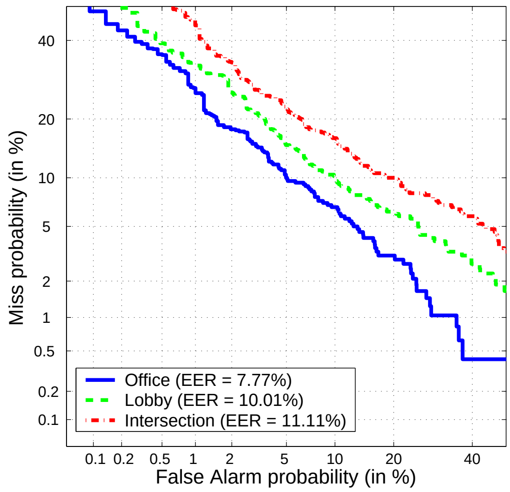

Figure 2: Detection-error tradeoff curves of multistyle trained models tested in three different locations.

crophone variability on speaker verification performance, we conducted experiments in which the system was trained from data collected with either the internal microphone or the earpiece microphone. Users enrolled by uttering five different name phrases three times each (once in each of the environment conditions) during the initial enrollment session. Subsequently, the trained system was then tested on data collected from the two different microphones.

Table 3 presents the results of the cross-microphone trials. From these results, it is observed that using different microphones during training and testing can also have a huge impact on system performance. In both cases, if the system was trained and tested using the same microphone, the EER was approximately 11%. However, in the two cross-condition cases, we see absolute performance degradations of over 7% and 11%. It should be noted that the earpiece microphone used in our experiments does suppress the high-frequency range where the speech signal-to-noise ratio is typically lower, while the internal microphone is a far-field microphone that typically yields lower signal-to-noise ratios especially in the high-frequency range. It is thus interesting to note that, similar to our cross-environment experiments, training with the noisier internal microphone data proved more robust in the cross-microphone tests than training with the earpiece microphone data.

4.5. Multistyle Training

Although more tedious for users, multistyle training (i.e. requiring a user to provide enrollment utterances in a variety of environments using a variety of microphones) can greatly improve robustness by creating diffuse models which cover a range of conditions. For our multistyle training experiments, the enrolled user recorded a single name phrase in each of the 6 testing conditions, essentially sampling all possible environment and microphone conditions. These models were then tested within each of the particular microphone or environment conditions. The results are shown in Figure 2. The multistyle models are far more resilient against performance degradations caused by changing environments than the models trained in single conditions (as evidenced by the obviously superior EERs of the multi-style models in Figure 2 compared with the cross-condition cases in Figure 1 and Table 2).

5. Conclusion

In this paper we have presented a new corpus developed for research aimed at the problem of robust speaker verification using mobile handheld devices with limited enrollment data. The corpus was collected in multiple noisy environments using different microphones. We have publically released this corpus for non-commercial research. It is available via download from our web site (http://groups.csail.mit.edu/sls/mdsvc).

In this paper, we have also presented preliminary experiments using the newly collected corpus. These experiments have shown that data collected in noisier environments tend to generate models that are more robust to mismatched environment conditions than data collected in quieter environments. Based on this observation we have begun studying methods for synthesizing multi-condition data from clean data and applying the parallel union model (an efficient missing feature method) to speaker models derived from the synthesized multi-condition data. This approach is known as *universal compensation* [13]. Experiments using this approach on our new mobile device corpus can be found in [14] and [15].

6. Acknowledgements

This work was sponsored in part by the Intel Corporation. The authors wish to thank Michael Deisher for his support of this project and his assistance with the Intel handheld device used in this work.

7. References

[1] P. Morin and J. Junqua, “A voice-centric multimodal user authentication system for fast and convenient physical access control,” *Proc. of Multimodal User Authentication Workshop*, pp. 19–24, 2003.

[2] G. Doddington, “Speaker recognition - identifying people by their voices,” *Proc. of IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[3] T. Hazen, E. Weinstein, and A. Park, “Towards robust person recognition on handheld devices using face and speaker identification technologies,” *Proc. of the International Conference on Multimodal Interfaces*, pp. 19–41, November 2003.

[4] T. Hazen, E. Weinstein, R. Kabir, A. Park, and B. Heisele, “Multi-modal face and speaker identification on a handheld device,” *Proc. of the Multimodal User Authentication Workshop*, pp. 113–120, December 2003.

[5] F. Bimbot *et al*, “An overview of the PICASSO project research activities in speaker verification for telephone applications,” *Proc. of Eurospeech*, September 1999.

[6] L. Lamel and J. Gauvain, “Speaker verification over the

telephone," *Speech Communication*, vol. 31, no. 2-3, pp. 141–154, 2000.

[7] C. Barras and J.L. Gauvain, "Feature and score normalization for speaker verification of cellular data," *Proc. of ICASSP*, April 2003.

[8] A. Park and T.J. Hazen, "ASR Dependent Techniques for Speaker Identification," *Proc. of ICSLP*, pp. 2521–2524, September 2002.

[9] A. Park and T. Hazen, "A Comparison of Normalization and Training Approaches for ASR-Dependent Speaker Identification," *Proc. of Interspeech*, October 2004.

[10] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, January 2000.

[11] R. Woo, "Exploration of small enrollment speaker verification on handheld devices," M.S. thesis, Massachusetts Institute of Technology, May 2005.

[12] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[13] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," *Proc. of ICASSP*, March 2005.

[14] J. Ming, T. Hazen, and J. Glass, "Speaker verification over handheld devices using realistic noisy speech data," *Proc. of ICASSP*, 2006.

[15] J. Ming, T. Hazen, and J. Glass, "A comparative study of methods for handheld speaker verification in realistic noisy conditions," *Proc. of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

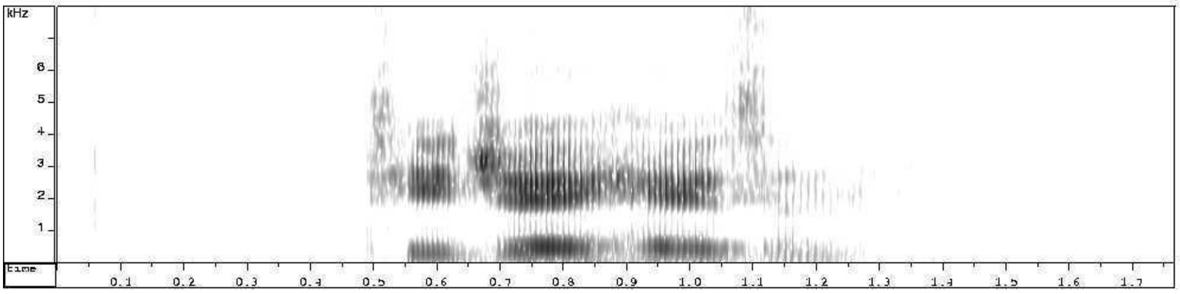

Figure 3: Spectrogram of sample phrase recorded in a quiet office with an earpiece microphone.

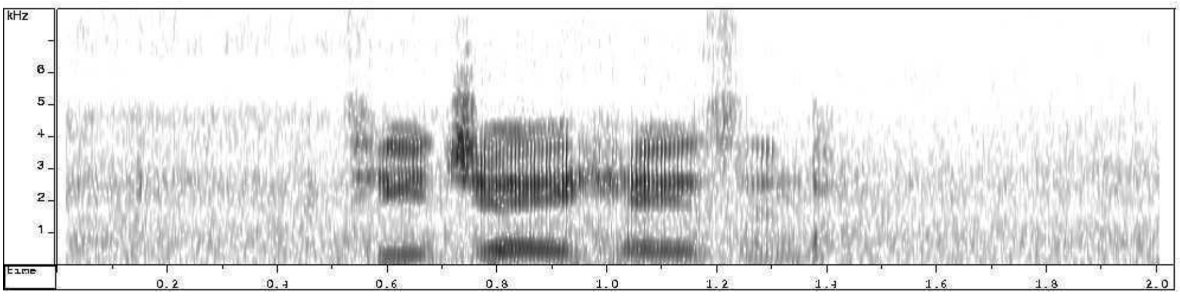

Figure 4: Spectrogram of sample phrase recorded at a noisy street corner with an earpiece microphone.

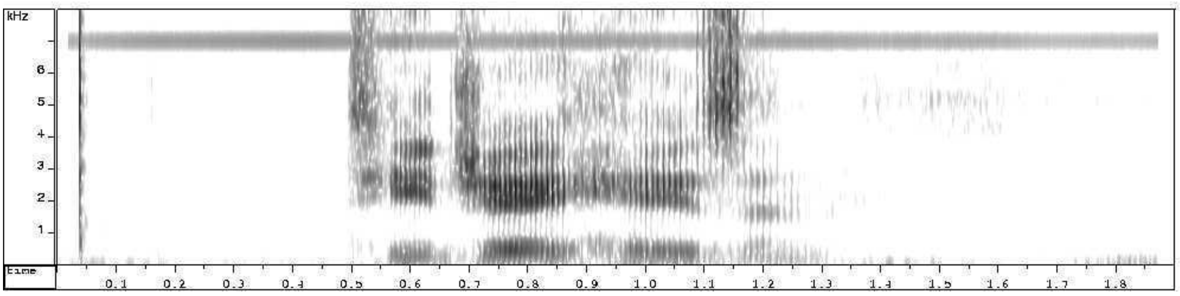

Figure 5: Spectrogram of sample phrase recorded in a quiet office with the handheld device's internal microphone.

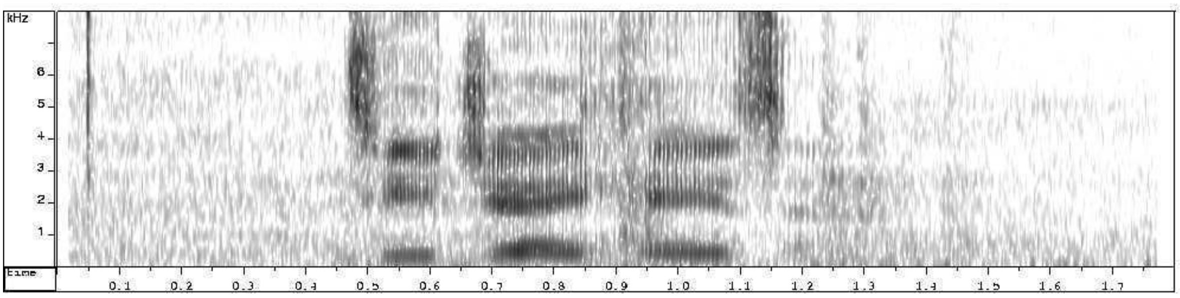

Figure 6: Spectrogram of sample phrase recorded at a noisy street corner with the handheld device's internal microphone.