SPOKEN COMMAND OF LARGE MOBILE ROBOTS IN OUTDOOR ENVIRONMENTS

Ekapol Chuangsuwanich, Scott Cyphers, James Glass, Seth Teller

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

{ekapolc, cyphers, glass, teller}@csail.mit.edu

ABSTRACT

We describe a speech system for commanding robots in human-occupied outdoor military supply depots. To operate in such environments, the robots must be as easy to interact with as are humans, i.e. they must reliably understand ordinary spoken instructions, such as orders to move supplies, as well as commands and warnings, spoken or shouted from distances of tens of meters. These design goals preclude close-talking microphones and "push-to-talk" buttons that are typically used to isolate commands from the sounds of vehicles, machinery and non-relevant speech.

We used multiple microphones to provide omnidirectional coverage. A novel voice activity detector was developed to detect speech and select the appropriate microphone to listen to. Finally, we developed a recognizer model that could successfully recognize commands when heard amidst other speech within a noisy environment. When evaluated on speech data in the field, this system performed significantly better than a more computationally intensive baseline system, reducing the effective false alarm rate by a factor of 40, while maintaining the same level of precision.

Index Terms— Human-robot interaction, real-time speech recognition, voice activity detection, modulation frequency.

1. INTRODUCTION

In order for a robot to function effectively in human environments, it must be able to react and respond to spoken input. This paper describes the development of methods for speech-enabling robotic platforms that operate in loosely organized outdoor warehouse environments and respond to spoken or shouted commands from supervisors or other workers. Our robotic platforms include an autonomous forklift for palletized cargo manipulation, a small rover for warehouse inventory listing, and a humanoid porter for handling boxes and other non-palletized items. In this paper, our focus will be on the forklift's speech recognition system, where the initial challenges have come from continuous listening for relevant distant speech in noisy environments. However, we have developed the system to facilitate porting to other generic work conditions and platforms.

Fig. 1: Supervisor speaking through a megaphone to command a robotic forklift (in blue box) 25 meters away.

Voice-enabled robots have existed in research labs for many years [1, 2]. Examples of areas of expertise for voice-enabled robots include receptionists [3], guides [4], and explorers [5], as well as indoor environments such as kitchens [6] and hospitals [7]. Within these areas some research has tried to extend the state-of-the-art to handle distant microphones [6] and more flexible dialogue strategies [8] that incorporate error correction [9], grounding [10], attention [11, 12], and even learning [13]. Due to the many challenges in achieving natural spoken human-robot interaction, researchers have usually constrained the problem in some way to focus on their particular research area of interest. For example, in much research on human-robot interaction, audio from the human is recorded via close-talking microphones. While in our project speech from a supervisor could be collected via a handheld internet device [14], this was not a constraint that we could impose on all humans in the warehouse environment. Thus, it was essential that the robots be "hearing"-enabled and be constantly listening for relevant speech input.

Another common constraint in human-robot interaction is to restrict the language to a set of limited phrases or a simple grammar that expresses alternatives. For our initial work, such a constraint was acceptable. In fact, our initial goal was to robustly detect shouting directed at the forklift as an

additional safety feature for halting the robot, especially the forklift, in a potentially hazardous situation. In this capacity, it was more important for us to detect any kind of shouted speech, so a grammar would have little value. Encouraged by our initial success, we have more recently augmented the role of continuous listening to enable nearby humans to issue a limited set of orders to the forklift, such as directing it to particular warehouse locations. Ultimately, we would like to expand these capabilities to allow for more sophisticated interactions, including clarifying dialogue.

The rest of this paper is organized as follows. Section 2 describes the design goals of our systems. The robotic forklift platform is introduced in Section 3. Section 4 provides an overview of the speech recognition system and its components, namely the voice activity detectors and the recognizer, which are further explained in Sections 5 and 6 respectively. Section 7 describes our testing environments, evaluates system performance, and points out possible future work. In Section 8 we provide some concluding remarks.

2. DESIGN CONSIDERATIONS

A number of elements of our system's design are dictated by our task, namely, outdoor warehouse management. The robotic platforms must be able to operate in existing human-occupied environments such as a military Supply Support Activity (SSA), our main deployment target for this research. The robots must operate outdoors on gravel and packed earth, which create different types of background noises. Other dominant background noises include engines, motors, construction, wind, beeping (from backward moving vehicles), and babble noise from existing personnel, making this a difficult environment for speech recognition.

The system also requires an effective command mechanism usable by military personnel with minimal training. We studied the language usage and general structure of warehouses in the SSA. A typical warehouse consists of three main zones: "receiving," "storage," and "issue." "Storage" is usually followed by letters from the NATO phonetic alphabet (Alpha for A, Bravo for B, etc.) specifying a particular storage area. The forklift is tasked with unloading pallets from the trucks in receiving, putting them into storage bays, or delivering the pallets to customers waiting in the issue zone. The humanoid porter is tasked with breaking down the pallets and distributing packages. The rover is tasked with warehouse area mapping and inventory listing. Table 1 shows some example of speech commands for the forklift. Figure 2 shows pictures of the 3 robots.

In order for personnel and pedestrians to operate safely around the robots, they must be able to continuously listen in noisy environments. The robots should be able to listen to commands spoken near the robots, shouted from several meters away, or even from 30 meters away in any direction via a megaphone. They also should be able to recognize shouted speech in emergency situations, describe their intentions, and respond to spoken commands in a transparent and predictable manner in order to be accepted in the work environment.

Type	Command
Summoning	Forklift come to Issue. Bot go to Receiving. Go to storage Alpha Charlie.
Pallet Manipulation	Forklift put this pallet in depot. Bot move the generator to issue. Pick up the inert ammo.
Safety Commands	Slow down forklift. Stop. Stop right now.

Table 1: Example speech commands for the robotic forklift.

3. ROBOTIC PLATFORM

The forklift platform is a Toyota 8FGU-15 manned lift truck, which is 106 cm wide, 208 cm long, and 223 cm tall. This is a large vehicle; it can cause acoustic shadowing depending on the location of the source. The forklift frame can also block some of the wind coming from the opposite direction. We chose to mount four Acoustic Magic Voice Tracker beamforming array microphones on the front, left, right, and rear side of the forklift to listen for speech on their respective sides of the forklift [15]. Due to the possible differences in quality caused by the large forklift size, each array is processed independently. The arrays are located 240 cm above the ground on the upper section of the forklift, to be as far from the engine as possible (Figure 2a). To display the forklift's intentions and responses to spoken commands, we added LED signage, marquee lights, and audio speakers to its chassis. The forklift operates with 4 quad-core laptops mounted in an equipment cabinet on the roof. However, the speech processing uses only a fraction of one CPU.

4. OVERALL SPEECH INTERACTION SYSTEM

The robots use a distributed publish-subscribe communications model [15]. Each of the microphones are independently sampled at 16KHz and periodically publish packets. The speech processing component subscribes to the audio for each of the microphones, processing each packet as it is received. Because the microphones publish their data as it becomes available, the processing of the audio data tends to be interleaved, so each packet can be processed as it arrives. If a packet is lost, silence is substituted for the missing data, since the robot can still listens from the other microphones.

The high-level system overview of the forklift's speech recognition system is shown in Figure 3. Separate voice activity detectors (VADs) listen to each microphone array channel. Since the speech signal from the microphone arrays contains

(a) Forklift (b) Porter (c) Rover

Fig. 2: The three robotic platforms. (a) The robot forklift. The front and the left microphone arrays are circled in red. (b) The robot porter. (c) The robot rover. The person in the back is holding a tablet which can be used for speech-based control.

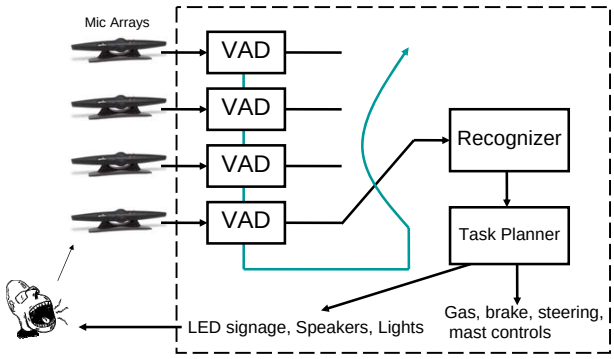

Fig. 3: Overview of the forklift's speech recognition system.

little high frequency content above 4KHz due to the long distances from the talker, we downsample the signal to 8KHz for subsequent processing. We also put a notch filter at 1.4KHz to filter out the forklift's beep. The VADs may trigger on multiple channels, for example if a person speaks from the left side of the forklift, the front, the rear, and the left side channels may trigger at the same time. We do channel selection by using scores from the VAD, instead of doing channel selection by the log-likelihood since some confusing words such as "stop" can give high likelihood scores [16].

When speech is detected on a channel by the VAD, it is pre-padded with 320 ms of waveform and forwarded to the recognizer. When the VAD detects the end of speech, 640 additional ms of signal are sent to the recognizer. These paddings are used to avoid clipping speech due to possible background noise. The recognizer is able to cope with the extra silence, but cannot recover from a clipped utterance.

The recognized commands are parsed. If a complete command is recognized, an appropriate message is published to the task planner. For most messages, the planner issues commands to the other modules on the forklift. After hearing a command, the task planner will repeat the command via the loudspeakers and LED signage, and then proceed with the command.

5. ROBUST VOICE ACTIVITY DETECTION

One of the key components for our system is a robust voice activity detector (VAD). A good VAD not only helps reduce the amount of computation required by the system, but it also helps increase the performance of the recognizer in terms of removing false alarms. In order to cope with our low SNR environments, we explored a two-stage system which uses a combination of two distinct features of speech, namely its harmonic spectral structure and rhythmic temporal structure.

5.1. Harmonicity

As illustrated in Figure 4, due to the possible large distances between the robot and a talker, non-vocalic portions of the recorded speech signal are often barely audible. Moreover, speech spoken through a megaphone loses high frequency components, such as those present in fricatives. Thus, detecting harmonicity structure for the detection of sonorants, provides a way to find candidate speech regions, even in low signal-to-noise ratio (SNR) environments. In this work, we compute harmonicity by using a simple periodicity that finds the size of the peak of the autocorrelation [17]. Since this measure is susceptible to periodic noise, (e.g., forklift beeps), we band-pass lifter in the cepstral domain prior to computing the autocorrelation. After the harmonicity features extract possible candidates for speech regions, the second part of the VAD, the modulation frequency, acts as an additional filter to help reduce possible false alarms.

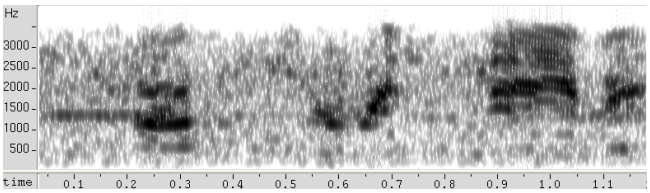

Fig. 4: Spectrogram of "Come to receiving" spoken through a megaphone.

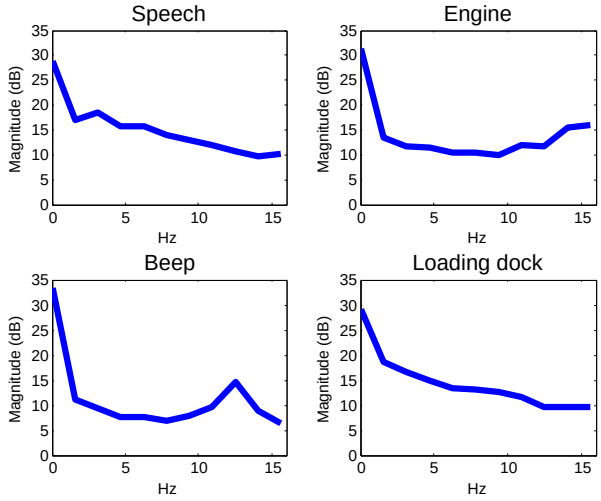

Fig. 5: MF comparison between speech and different type of noises. Top left: Speech. Top right: Engine noise. Bottom left: Beep noise. Bottom right: Loading Dock noise.

5.2. Modulation Frequency

For the rhythmic structure of speech, we extract Modulation Frequencies (MFs) from multiple sub-bands in the range of 160-2400Hz. The MF is the frequency that is modulating each sub-band. Recently, MFs have been receiving attention from the robust speech recognition community [18]. MFs have been shown to have potential as a robust feature for speech/non-speech classification [19]. The MF of speech usually has a peak around 4 Hz, which is the syllable rate of human speech. The MF spectrum also has a slower decay than many kinds of noise, as shown in Figure 5. Drullman et al. have shown that only the low MFs, ranging from 0 to 16 Hz, are important for human language understanding [20]. We fit Legendre Polynomials to extract the shape of the MF up to 16 Hz. The polynomial coefficients are used to classify speech/non-speech frames via Support Vector Machines.

We evaluated the effectiveness of our VAD on clean speech data with digitally added noise. Clean speech was recorded from 23 speakers shouting 25 stop commands. We added street, loading dock, babble, beep, wind, and engine noise at various SNR values ranging from -5 to 15 dB. We compared our VAD system with another robust feature, namely, Relative Spectral Entropy (RSE) [21]. The ROC curves in Figure 6 show that our VAD outperforms RSE significantly. The average Equal Error Rate (EER) over all noise conditions for our VAD is 3.6%, while RSE's EER is 6.7%. Note that for this data set, a standard VAD such as the one in G.729B [22] operates at 0.05-0.1 miss detection rate and 40-80 false alarms/minute depending on the noise type.

5.3. Channel Selection

The score used to select the channel to send to the recognizer is the harmonicity value. Since the harmonicity is the autocorrelation of the input signal, it can be considered as a crude estimate of the SNR [17]. We tested our channel selection method by speaking to the forklift from various directions while it performed various tasks. In 412 trials, the harmonicity selection method chose the closest microphone 84% of the time. Note that the closest microphone does not necessarily have the best SNR, as wind noise changes depending on the wind direction. By using the harmonicity value, which is already computed, we reduce the amount of computation required while maintaining reasonable performance.

6. THE RECOGNIZER

Automatic speech recognition is performed using our small footprint landmark-based speech recognizer [23]. Our initial effort used a context-free grammar to represent possible spoken commands for this task. There were a total of 57 command words in the vocabulary. Since we expected some of the detected speech to be out of domain (OOD) (i.e., not directed at the robot), we incorporate an explicit OOD command that is modeled by a single Gaussian mixture model trained on generic speech. No explicit noise models were trained for this grammar; all noises were modeled by a silence model.

The acoustic model was adapted from a telephone-based model using three sources of data collected from the array microphones. The first source of data was recorded in an indoor hanger environment and included over 3,600 utterances of stop commands from 18 talkers under different acoustic conditions (quiet, motor noise, babble noise, beep noise). The second source of data was recorded in outdoor paved and gravel parking lot environments, and consisted of nearly 1,100 shouted commands from 16 talkers. The third source of data consisted of 70 utterances from 5 talkers issuing commands through a megaphone in an actual SSA environment.

7. EVALUATION AND DISCUSSION

The evaluation of the robot speech processing system was based on data that was collected in a real SSA outdoor warehouse in Fort Lee, Virginia. Some of these data were collected during a series of live demonstrations that illustrated the capabilities of the robots. During the data collection there was nearby construction noise and noise from the PA system giving explanations to the audience. The spoken commands

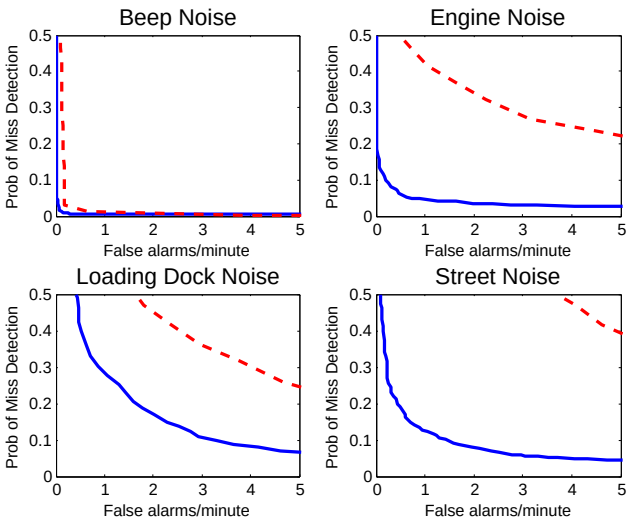

Fig. 6: ROC of the VADs. Top left: Beep noise. Top right: Engine noise. Bottom left: Loading dock noise. Bottom right: Street noise. Our VAD system in blue. RSE in red dashed line. The x-axis shows the probability that a 320 ms speech frame will be misclassified as non-speech. The y-axis shows the number of false alarm frames per minute.

were given to the forklift via a megaphone from 30 meters away. The evaluation data was also augmented with speech data that was collected at the SSA during earlier trial runs, consisting of shouts from several meters away and additional megaphone speech trials. The total amount of evaluation data amounted to 130 minutes of recorded data. During this time, there were 19 shouted commands at SNR values ranging from 10 to 25 dB. The commands were relatively sparse due to the nature of the interactions between the humans and the robotic forklift (i.e., the forklift had to execute the command).

The evaluation metric was based on correct detection and understanding of the spoken command, rather than word-level recognition accuracy. For example, "Move this pallet to storage Alpha" and "Put the pallet in storage area Alpha" are considered the same even though the exact words are different. Errors were categorized into three different types. A spoken command that was recognized as OOD or that the VAD failed to detect was considered a miss. A command that was detected due to a false VAD trigger of non-speech or misrecognition of OOD speech was considered a false alarm. Finally, a correctly detected but ultimately misunderstood command was considered an error.

We compared this speech detection and recognition framework with an earlier baseline system that was used purely as a safety feature to detect only the command "Forklift stop." The baseline system fed overlapping 2 second chunks of speech into dual speech recognizers/channel, for a total of eight recognizers for the four microphone arrays [15]. For evaluation, we augmented the baseline configuration with

System	Correct	Miss	Error	False Alarm
Current	13 (72%)	2 (11%)	4 (22%)	11 [0.8]
Baseline	13 (72%)	2 (11%)	4 (22%)	425 [32.7]

Table 2: Performance on demonstration data. The numbers outside the brackets correspond to counts for each type of occurrence. The numbers in brackets correspond to the percentage over the total number of commands. Numbers in square brackets correspond to false alarms per 10 minutes.

the new speech recognition acoustic and language models. Thus, we were mainly evaluating the effectiveness of the VAD component to reduce computation and false alarms. Note that the multiple speech recognizers of the baseline system could produce conflicting commands; in these cases we selected the correct output if it was available (i.e., an upper bound).

As shown in Table 2, the current system was able to achieve a false alarm rate of less than 1 false alarm every 10 minutes. This indicates that the VAD system is able to filter out most of the non-speech portions of the audio. Moreover, the current system achieves the same level of performance as the baseline system in terms of understanding, which indicates that the channel selection method, even with less computation required, does not degrade performance. On the current system, all of the false alarms are stop commands. This is due to our design choice to accept the easily confusable single word "stop" as a possible command. However, this command is required as it is what naturally comes to mind in human-human interaction! The current system was able to detect a real emergency shout "stop stop stop" from one of our team members directed at another team member holding the emergency stop button near the forklift during one of our trial runs. This indicates the potential for the system to be able to cope with agitated speech in the future.

Although there are a significant number of misses and errors, these were mostly due to insufficient loudness. Such errors often occur in human-human interaction across distances (as depicted in Figure 1). However, after a miss or error happens, the speaker usually repeats the command more loudly, which makes the forklift able to correctly recognize the latter tries. This behavior is consistent with a human working with another human forklift operator in noisy environments, another feature that is helpful in integrating the robots into the work environment without forcing humans to change their behavior.

Several parts of the system can be improved. We are currently working on ways to reduce the effects of background noise so as to improve recognition accuracy. The robot's vocabulary and grammar should be easily extensible to new environments and task domains. The robot should support some sort of supervisor-authentication mechanism, perhaps through speaker recognition. Reduction of false alarms, and interpre-

tation of deictic gestures, would be facilitated by increased integration of the robot's speech understanding module and its vision- and lidar-based situational awareness module.

8. CONCLUSION

This paper described our speech recognition system for distant speech in robots designed for operating in human-occupied outdoor military warehouse. Creating a safe and reliable robot requires omnidirectional continuous listening in noisy environments while keeping computation costs low. To accomplish this, we introduced the use of multiple microphones combined with a novel voice activity detector and channel selection method. Live testing at a military SSA has shown that our system was able to interact reliably with humans in the presence of noise.

9. REFERENCES

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," 2002.

[2] R. Prasad, H. Saruwatari, and K. Shikano, *Advanced Robotics*, vol. 18, pp. 533–564, 2004.

[3] R. Nisimura, T. Uchida, A. Lee, et al., "ASKA: receptionist robot with speech dialogue system," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2002, vol. 2, pp. 1314–1319.

[4] W. Burgard, A. Cremers, D. Fox, et al., "The interactive museum tour-guide robot," 1998, pp. 11–18.

[5] O. Lemon, A. Bracy, E. Gruenstein, and S. Peters, "A multi-modal dialogue system for human-robot conversation," in *Proc. NAACL*, 2001.

[6] R. Stiefelhagen, C. Fügen, P. Gieselmann, et al., "Natural human-robot interaction using speech, head pose and gestures," in *Meetings with Omnidirectional Cameras, International workshop on Multimedia Technologies in E-learning and Collaboration*, 2004.

[7] D. Spiliotopoulos, I. Androutsopoulos, and C. Spyropoulos, "Human-robot interaction based on spoken natural language dialogue," in *Proc. European Workshop on Service and Humanoid Robots*, 2001, pp. 25–27.

[8] J. Bos, E. Klein, and T. Oka, "Meaningful conversation with a mobile robot," in *Proc. 10th Conf. of the European Chapter of the Association for Computational Linguistics*, 2003, pp. 71–74.

[9] H. Holzapfel and P. Gieselmann, "A way out of dead end situations in dialogue systems for human-robot interaction," in *Proc. IEEE/RAS Int. Conf. on Humanoid Robots*, 10-12 2004, vol. 1, pp. 184–195.

[10] P. Gieselmann and A. Waibel, "What makes human-robot dialogues struggle?," in *Proc. Semantics and Pragmatics of Dialogue Workshop*, 2005.

[11] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," 2002.

[12] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: joint attention for human-robot interaction," in *Proc. 10th IEEE Int. Workshop on Robot and Human Interactive Communication*, 2001, pp. 512–517.

[13] L. Seabra Lopes and A. Teixeira, "Human-robot interaction through spoken language dialogue," 2000.

[14] A. Correa, M. Walter, L. Fletcher, et al., "Multimodal interaction with an autonomous forklift," in *5th ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2010, pp. 243–250.

[15] S. Teller, M. Walter, et al., "A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2010.

[16] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proc. ICASSP*, 2000, vol. 3, pp. 1747–1750.

[17] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. the Institute of Phonetic Sciences*, 1993, pp. 97–110.

[18] S. Ganapathy, S. Thomas, and H. Hermansky, "Comparison of modulation features for phoneme recognition," 2010, pp. 5038 –5041.

[19] H. You and A. Alwa, "Temporal modulation processing of speech signals for noise robust ASR," in *Proc. Interspeech*, 2009, pp. 36–39.

[20] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *JASA*, vol. 95, pp. 1053–1064, 1994.

[21] A. Ouzounov, "Robust features for speech detection - a comparative study," in *Int. Conf. on Computer Systems and Technologies*, 2005, pp. 19/1–19/6.

[22] A. Benyassine, E. Shlomot, et al., "ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *Communications Magazine, IEEE*, vol. 35, pp. 64 –73, Sept 1997.

[23] I. Hetherington, "PocketSUMMIT: Small-footprint continuous speech recognition," in *Proc. Interspeech*, 2007.