A Distributed System for Recognizing Home Automation Commands and Distress Calls in the Italian Language

Emanuele Principi[1], *Stefano Squartini*[1], *Francesco Piazza*[1], *Danilo Fuselli*[2], *Maurizio Bonifazi*[2]

[1]Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
[2]FBT Elettronica Spa, Recanati (MC), Italy
{e.principi,s.squartini,f.piazza}@univpm.it, {danilo.fuselli,maurizio.bonifazi}@fbt.it

Abstract

This paper describes a system for recognizing distress calls and home automation voice commands in a smart-home. Distress calls are recognized with the purpose of assisting people in their own homes: when they are detected, a phone call is automatically established with a contact in a address book and the person can request for assistance. The voice call is established through a voice over ip stack, with hands-free communication guaranteed by an acoustic echo canceller. The acoustic environment is constantly monitored by several low-consuming devices distributed throughout the home. In each device, a voice activity detector detects speech segments, and a speech recognition engine recognizes commands and distress calls. Robustness to environmental disturbances has been increased by employing Power Normalized Cepstral Coefficients and by using an adaptive algorithm for interference cancellation. An Italian speech corpus of home automation commands and distress calls has been developed for evaluation purposes. The corpus has been recorded in a real room using multiple microphones, and each sentence has been uttered both in normal and shouted speaking styles. The system performance has been assessed in terms of commands/distress recognition accuracy in order to prove the effectiveness of the approach.

Index Terms: automatic speech recognition, ambient assisted living, home automation, emergency state recognition

1. Introduction

The ever increasing percentage of old people in the most developed countries is posing a great challenge in social healthcare systems [1]. The effort required by formal care givers for supporting older people can be enormous, and this requires an increase in the efficiency and effectiveness of today's care. One way to achieve such a goal is the use of technologies for supporting and assisting people in their own homes.

Usually, these technologies incorporate application dependent sensors, such as vital sensors or video cameras, but recent studies [2, 3] have demonstrated that people prefer less invasive sensors, such as microphones. Several works appeared in the literature that exploit audio signals only: in [2], the presented system acquires audio signals using multiple microphones positioned on the ceiling and on floor lamps and emergencies are detected by classifying audio events (e.g., clapping, coughing). Based on the classification outcomes, a reasoning model then identifies dangerous situations and determines if an alarm should be raised or not.

Vacher *et al.* [4, 5] describe a system for telemonitoring based on speech recognition technology. The system is composed of an audio analysis module which segments the incoming signal, and then classifies it as being speech or not. In the first case, the segment is processed by a speech recognizer that captures home automation commands as well as distress calls. In the second, a sound classifier determines the class of the signal (e.g., door slap, step, object falling, etc.).

In [6], keywords recognition and voice stress classification are combined in a single system. The keywords recognizer, based on Dynamic Time Warping and k-Nearest Neighbour, detects words that indicate emergency situations. The voice stress classifier determines if the speech on input is affected by fear or anger. The two outputs are then used by an action-taking module that undertakes the appropriate actions.

The system presented in this paper integrates the automatic recognition of emergency states and home automation commands with remote assistance and hands-free communication. The acoustic environment is constantly monitored to detect speech signals by means of a Voice Activity Detector (VAD), and a speech recognizer based on PocketSphinx [7] detects distress calls and voice commands. Robustness against noise and reverberation is increased by integrating Power Normalized Cepstral Coefficients (PNCC) [8] in the engine and reducing known sources of interferences by means of an interference cancellation module. When a distress call is detected, the system automatically establishes a hands-free communication with one of the contacts present in a address book. The person can then ask for assistance, explain the reason of the distress call and be reassured. The paper describes the entire architecture of the proposed system and presents its current state of development. Being an ongoing project, the preliminary evaluation assesses the performance of the recognition system, while future works will be devoted to subjective testing of the complete system. The performance of the system have been assessed by building a new corpus of Italian speech acquired in a realistic scenario.

The outline of the paper is the following: Section 2 presents the system architecture, as well as the algorithms for commands and distress calls recognition, and hands-free communication. Section 3 presents ITAAL, a new Italian corpus of home automation commands and distress calls. Section 4 describes the experiments conducted to assess the recognition performance. Finally, Section 5 concludes the paper and presents future developments.

2. The proposed system

The proposed system integrates acoustic monitoring for emergency detection, hands-free communication services and recog-

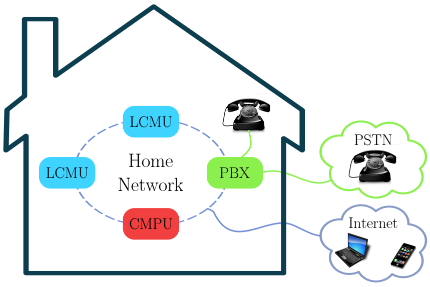

Figure 1: Architecture of the proposed system.

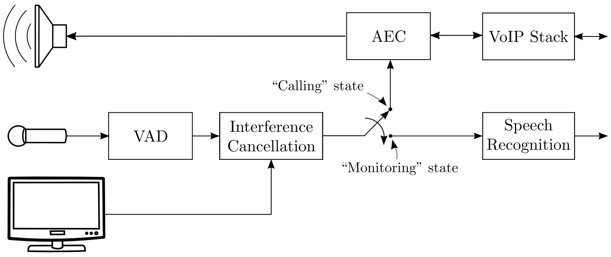

Figure 2: The Local Multimedia Control Unit.

nition of voice commands for controlling home automation services. The two functional units that compose the system are the Local Multimedia Control Unit (LMCU), and the Central Management and Processing Unit (CMPU) (Figure 1). The LMCU is equipped with a touch screen LCD monitor, a video camera, a microphone and a loudspeaker. Users can control the system through the touch screen interface or through vocal commands. The unit can also reproduce multimedia contents stored in the CMPU. Emergency states are detected by recognizing distress calls, and hands-free communication services are supported by means of a Voice over IP (VoIP) stack. Depending on the dimension of the building and on the monitoring requirements, several LMCUs can compose the system. For example, one LMCU can be present in each room of a house. The CMPU integrates home and building automation services, audio content delivery services, and management of emergency states.

The LMCUs and the CMPU are connected to the same local area network (either wired or wireless), and communicate by means of the TCP/IP protocol. Besides the LMCUs and the CMPU, Figure 1 also shows a Private Branch eXchange (PBX) that is connected to the same network and that interfaces the system with the Public Switched Telephone Network (PSTN) and analogue phone devices. The PBX can be a separate module or it can be a component of the CMPU.

Each LMCU constantly monitors the acoustic environment to detect both home automation commands (e.g., "switch on the light"), and distress calls (e.g., "help!"). In the latter case, the system automatically calls one or more contacts present in a shared address book. For each contact, the user can specify a calling priority in case of emergency. When a home automation command is detected, the LMCU directly performs the requested action, or it sends the command to the CMPU that provides for its execution.

2.1. Local multimedia control unit

The block-scheme of the local multimedia control unit is shown in Figure 2. The audio signal is acquired by means of a single microphone, then a VAD selects the audio segments containing the voice signal. In the current implementation, speech activity detection follows a simple energy-based approach [9], but future works will take into consideration more advanced approaches [10]. The interference cancellation module reduces sounds coming from known audio sources (e.g., from a television or a radio). In the current implementation, the interference is acquired through an analogue line-in interface.

The LMCU operates in two states: "monitoring", and "calling". In the first, the speech recognition engine is active and detects home automation commands and distress calls. The "calling" state occurs when a phone call is ongoing: in this case, the speech recognition engine is disabled and the Acoustic Echo Canceller (AEC) and the entire VoIP stack are active. The LMCU enters in this state either when it receives a phone call, or when the call is automatically started because of the detection an emergency. Hands-free communication is guaranteed by the VoIP a stack and the AEC. In brief, the VoIP stack is based on the Session Initiation Protocol (SIP) [11], the Real-time Transport Protocol (RTP) and RTP Control Protocol. The hardware platform of the LMCU is based on the ARM Cortex-A8 CPU and on the embedded Linux distribution by Linaro[1].

2.2. Acoustic Echo & Interference Cancellation

In hands-free communication, the acoustic coupling between the loudspeaker and the microphone via the impulse response of the room in which they are located creates acoustic echoes. Acoustic echo cancellation algorithms are needed in order to remove the echo and to ensure that the ongoing communication is of sufficient quality. Several approaches have been proposed in the literature to address the AEC problem [12]. In the system presented in this paper, the approach implemented in the Speex codec has been used [13, 14]. Without entering into the details, AEC in Speex is based on the Multidelay Block Frequency Domain (MDF) adaptive filter, and double-talk is handled by dynamically varying the filters coefficients adaptation step-size.

The same algorithm has been employed to reduce a known source of interference, such as the sound coming from a television or a radio. As aforementioned, the interference is acquired connecting the output of the source device (e.g., television, radio, etc.), to the input of the LMCU. This signal then becomes the "far-end" (reference) signal of acoustic echo cancellation algorithms. Testing of the algorithm in synthetic conditions reported that convergence is reached after about 5 s, with an Echo Return Loss Enhancement (ERLE) of 30 dB and a CPU occupancy of 6.65%. In the experiment, the filter length was set to 1024 taps, the sampling frequency to 16 kHz and the synthetic impulse response was 1024 taps long with a reverberation time of 250 ms.

2.3. Commands and distress calls recognition

Commands and distress calls recognition is performed by means of the PocketSphinx [7] speech recognition engine. PocketSphinx has been chosen because it includes several optimizations that make it suitable for embedded devices. The next sections illustrate the feature extraction stage, and the acoustic and the language models structures and parameters.

[1] http://www.linaro.org

2.3.1. *Feature extraction*

Currently, in most speech recognition systems features are represented by Mel-Frequency Cepstral Coefficients (MFCC) and their first and second derivatives [15]. Cepstral mean normalization is usually applied on the static coefficients to improve the robustness against channel mismatch.

In order to increase the robustness against noise and reverberation, several approaches have been proposed that operate before [16–18], or directly inside the MFCC feature extraction pipeline [18–21]. An alternative solution is using a different set of features that are intrinsically more robust. Power Normalized Cepstral Coefficients (PNCC) [8] are member of this family of approaches and they have demonstrated their effectiveness at the cost of a modest increment of computational burden. The main innovations with respect to MFCCs are the replacement of the logarithmic non-linearity with a power function non-linearity, the introduction of the "medium-time processing" stage that operates on segments with duration of 50-120 ms, the use "asymmetric non-linear filtering" to estimate background noise and the use of "temporal masking".

In the proposed system, features are extracted from signals sampled at 16 kHz and PNCC are parametrised as in [8]. The final feature vector is composed of 13 mean normalized static coefficients and their first and second derivatives.

The PNCC feature extraction pipeline has been implemented in C language[2], and integrated in PocketSphinx.

2.3.2. *Acoustic model*

The acoustic model has been trained and parametrised using the APASCI corpus [22]. The corpus is composed of Italian speech utterances recorded in a quiet room with a vocabulary of about 3000 words. Training has been performed on the training set part of the speaker independent subset. This set is composed of 1310 sentences uttered by 30 females and 30 males having a total duration of about 105 minutes. The APASCI test set is composed of 860 utterances spoken by 20 males and 20 females and it has a total duration of about 69 minutes.

The acoustic model structure is continuous with 3 states per phones (without skip) and 200 senones (tied states). The number of gaussians per state has been set to 4. These values have been selected using the APASCI speaker independent test set as development set and a world loop grammar as language model. The obtained word recognition accuracy is 64.5%. Special care must be taken to reject out of grammar sentences in order to avoid possible false positives. The current version of the PocketSphinx (version 0.8) is not able to perform rejection using generic word models [23] or with garbage models[3] as in keyword spotting [24]. In addition, the "confidence score" associated to each result is reliable only for vocabularies of more than 100 words, a number exceeding the vocabulary size of the application scenario taken into consideration.

The approach to reject of out of grammar words is inspired by the technique of Vertanen [25]. This consists in training a special "garbage phone" that represents a generic word phone replacing the actual phonetic transcription of a percentage of the vocabulary words with a sequence of garbage phones. Then, in the recognition dictionary a garbage word is introduced whose phonetic transcription is the single garbage phone. Here, training of the garbage phone has been performed substituting 10% of the phonetic transcriptions of the words present in the training vocabulary. Words have been chosen so that each phone of the Italian language is equally represented, so that they are all trained with the same amount of data. The CPU occupancy of the recognition engine (comprising the feature extraction stage) is about 20%.

```
public <SENTENCE> = <COMMAND> | <DISTRESS_CALL> | GARBAGE;
<COMMAND>         = ((accendi|spegni) la luce) |
                    ((alza|abbassa) la temperatura);
<DISTRESS_CALL>   = aiuto | ambulanza;
```

Figure 3: An excerpt of the recognition grammar.

Robustness to mismatches between training and testing conditions can be improved by adapting the acoustic model. Before using the system for the first time, the user is asked to speak a set of phonetically rich sentences so that all phones of the Italian language are trained using the same amount of data. Adaptation is performed by means of Maximum Likelihood Linear Regression (MLLR) [26], since it works best when the amount of data for adaptation is limited.

2.3.3. *Language model*

The language model is represented by a simple finite state grammar (an excerpt is shown in Figure 3). The grammar comprises both home automation commands and distress calls. The word `GARBAGE` is mapped to the GP phone in the dictionary, and sequences of one or more garbage words are automatically ignored. Stop-words, such as articles and prepositions, are also ignored by the system.

3. The ITAAL speech corpus

In order to assess the recognition performance of the system, a speech corpus of home automation commands and distress calls has been created[4]. The corpus is composed of utterances spoken by 20 native Italian speakers, half males and half females. The average age of the speakers is 41.70 years with a standard deviation of 11.17 years. Recordings have been performed using both a headset microphone (AKG C 555 L) and an array composed of four C 400 BL hypercardiod microphones spaced by 4 cm and placed on a table 80 cm height. The room measured 9.7 m×8.0 m×2.9 m, with a reverberation time (T_{60}) of 0.72 s. Each person spoke the corpus sentences standing in front of the microphone array at a distance of 3 m. Signals have been acquired with a sample rate of 48 kHz using a MOTU 8pre sound interface, and they were later downsampled to 16 kHz. People were asked to read three groups of sentences in Italian: home automation commands (e.g., "close the door"), distress calls (e.g., "help!", "ambulance!") and phonetically rich sentences. The latter sentences have been extracted from the "speaker independent" set of the APASCI corpus and they cover all the phones of the Italian language.

Every sentence was spoken both in normal and shouted conditions, with the home automation commands and the phonetically rich sentences pronounced without emotional inflection. In contrast, people were asked to speak distress calls as they were frightened. Figure 4 on the left shows the pitch distribution for the 10 male speakers both for the normal and shouted speaking style utterances. Females' pitch distribution presents the same behaviour, but it is not reported for the sake of conciseness. It can be noticed that in shouted utterances, the pitch shifts towards higher frequencies and the overall variance is increased.

[2] http://www.a3lab.dii.univpm.it/projects/libpncc

[3] http://cmusphinx.sourceforge.net/wiki/faq

[4] http://www.a3lab.dii.univpm.it/projects/itaal

	Commands	**Distress**	**APASCI**
Number of sentences	15	5	6
Total words	45	8	52
Unique Words	18	6	51
Duration (minutes)	39	13	17
Repetitions	3	3	1

Table 1: Details on the composition of ITAAL.

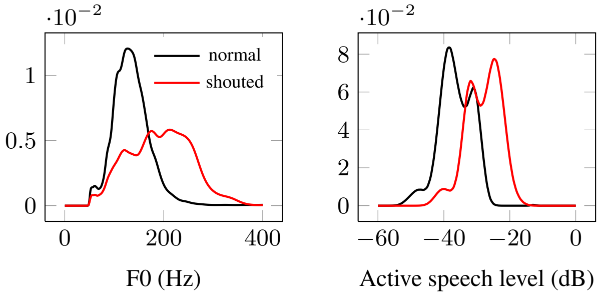

Figure 4: Left: the pitch distribution for the 10 male speakers. Right: the active speech level for the 20 speakers.

Figure 4 on the right shows the distribution of the active speech level measured as described in [27]. Again, it is evident the difference between normal and shouted speaking style, with the latter presenting a higher level. This is consistent with other studies on shouted speech, such as [28]. Table 1 shows the details of the corpus: note that the reported durations are related to the active speech segments as measured following [27]. The average signal-to-noise ratio of the headset microphone signals is 51.46 dB, while the distant microphone signals one is 34.08 dB.

4. Experiments

This section describes the experiments conducted to assess the performance of the recognition module parametrised as described in Section 2.3. Experiments have been performed on the ITAAL corpus, in particular on the signals acquired from headset microphone and from one of the central microphones of the array. Based on the Italian speech corpus composition, the system has been evaluated on four tasks: recognition of home automation commands and distress calls uttered in normal and shouted speaking styles. Each task has been evaluated with and without adapting the acoustic model.

For each task, the evaluation metric is the sentence recognition accuracy, i.e. the ratio between the number of correctly recognized sentences and the total number of sentences.

4.1. Performance without speaker adaptation

Table 2 (“Baseline” rows) shows the obtained results. The accuracy on the headset microphone signals exceeds 90% in each task, with the only exception being shouted commands recognition. In shouted speech, a similar performance decrease can be also observed in distress call recognition. The results obtained using the single distant-talk microphone can certainly be improved, in particular in the command recognition task. The main reason is the high mismatch between training and testing conditions, with test files being highly reverberated (as pointed out in Section 3, the T_{60} of the room is 0.72 s). Consider also that the acoustic model parameters have been chosen using the APASCI test set, thus on noise and reverberation free signals (see Section 2.3.2). The performance difference between normal and shouted conditions is evident also on the distant microphone signals, and this is consistent with the results in [29].

	Commands		**Distress calls**		**Average**
Headset	Normal	Shouted	Normal	Shouted	
Baseline	94.22	87.89	99.67	93.00	93.70
MLLR	95.77	96.11	100.00	95.33	96.80
Distant					
Baseline	19.13	10.67	70.67	45.33	36.45
MLLR	37.15	37.67	85.00	72.67	58.12

Table 2: Commands and distress calls recognition accuracy (%).

4.2. Performance with speaker adaptation

Acoustic model adaptation has been performed separately for each speaker, for each talking style (normal, shouted) and for each microphone (headset, distant) on the phonetically rich sentences of ITAAL.

Table 2 (“MLLR” rows) shows the obtained results: on the headset microphone, the introduction of speaker adaptation produces a significant performance improvement over non-adapted results. Other than for the acoustic conditions, the improvement can be attributed also to the different accent between the APASCI corpus speakers (northern Italy) and the ITAAL ones (central Italy). As expected, shouted speaking style improvements are more significant.

The improvement of recognition accuracy using the distant talking microphone is more evident, reaching an average of 21.67%. It is interesting to note that while commands recognition performance can certainly be improved, distress calls recognition can still be considered satisfactory for being employed in a real scenario.

5. Conclusions

In this paper, a system for the emergency detection and the remote assistance in a smart-home has been presented. The proposed system is composed of multiple local multimedia control units that actively monitor the acoustic environment, and one central management and processing unit that coordinates the system. The acoustic environment is monitored by means of a voice activity detector, an interference removal module, and a speech recognizer based on the PocketSphinx open source engine recognizes distress calls and voice commands addressed to the home automation system. When a distress call is detected, a phone call is automatically established with one of the contacts of the address book. Hands-free communication is possible by means of a VoIP stack based on the SIP protocol, and the Speex acoustic echo canceller. Experiments have been conducted to assess the recognition capabilities using ITAAL, a new Italian speech corpus of distress calls and home automation commands recorded in a realistic scenario. The results showed that adapting the recognizer acoustic model by means of phonetically balanced sentences the system is able to achieve a distress call recognition accuracy of 85.00% in normal speaking styles utterances, and 72.67% in shouted speaking style ones.

In future works, algorithms will be integrated to improve the recognition accuracy on shouted speech, for example detecting the speaking style as in [30, 31], and multiple microphone channels will be employed to increase the performance using distant talking conditions. Finally, the interference cancellation algorithm will be tested in real scenarios and a subjective evaluation will be carried out to establish the effectiveness of the overall system.

6. References

[1] K. Giannakouris, "Aging characterizes the demographic perspectives of the European societies," 2008, Eurostat 72.

[2] S. Goetze, J. Schroder, and S. Gerlach, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[3] M. Ziefle, C. Rocker, and A. Holzinger, "Medical technology in smart homes: Exploring the user's perspective on privacy, intimacy and trust," in *Proc. of COMPSACW*, Jul. 18-22 2011, pp. 410–415.

[4] M. Vacher, A. Fleury, J.-F. Serignat, N. Noury, and H. Glasson, "Preliminary evaluation of speech/sound recognition for telemedice application in a real environment," in *Proc. of Interspeech*, Brisbane, Australia, Sep. 2008, pp. 496–499.

[5] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment," in *Proc. of Eusipco*, Bucharest, Romania, Aug. 27-31 2012, pp. 1663–1667.

[6] K. Drossos, A. Floros, K. Agavanakis, N.-A. Tatlas, and N.-G. Kanellopoulos, "Emergency voice/stress-level combined recognition for intelligent house applications," in *Proc. of the 132nd AES Convention*, Budapest, Hungary, Apr. 26-29 2012, pp. 1–11.

[7] D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, and A. Rudnicky, "PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. of ICASSP*, vol. 1, Toulouse, France, May 15-19 2006, pp. 185–188.

[8] C. Kim and R. M. Stern, "Power-normalized coefficients (PNCC) for robust speech recognition," in *Proc. of ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4101–4104.

[9] A. M. Peinado and J. C. Segura, *Speech Recognition Over Digital Channels: Robustness and Standards*. West Sussex, England: John Wiley & Sons, Ltd, 2006.

[10] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-Life Voice Activity Detection with LSTM Recurrent Neural Networks and application to Hollywood Movies," in *Proc. of ICASSP*, 2013, to appear.

[11] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, *SIP: Session Initiation Protocol*, RFC 3261, IETF Std., 2002.

[12] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, and S. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York: Springer, 2001.

[13] J.-M. Valin, "Speex: A free codec for free speech," in *Proc. Linux Conf.*, Australia, 2006.

[14] J.-M. Valin, "On Adjusting the Learning Rate in Frequency Domain Echo Cancellation With Double-Talk," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 3, pp. 1030–1034, 2007.

[15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, 1980, pp. 357–366.

[16] E. Principi, S. Cifani, R. Rotili, S. Squartini, and F. Piazza, "Comparative evaluation of single-channel MMSE-based noise reduction schemes for speech recognition," *Journal of Electrical and Computer Engineering*, p. 6, 2010, Article ID 962103.

[17] R. Rotili, E. Principi, S. Squartini, and B. Schuller, "A real-time speech enhancement framework in noisy and reverberated acoustic scenarios," *Cognitive Computation*, pp. 1–13, 2012.

[18] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, p. 17, 2009, Article ID 942617.

[19] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental robust speech and speaker recognition through multi-channel histogram equalization," *Neurocomputing*, vol. 78, no. 1, pp. 111–120, 2012.

[20] V. Stouten, "Robust Automatic Speech Recognition in Time-varying Environments," Ph.D. dissertation, K. U. Leuven, Leuven, the Netherlands, 2006.

[21] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 5, pp. 1061–1070, Jul. 2008.

[22] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Automatic segmentation and labeling of english and italian speech databases," in *Proc. of Eurospeech*, Berlin, Germany, Sep. 22-25 1993, pp. 653–656.

[23] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in *Proc. of ICSLP*, vol. 1, Beijing, China, 2000, pp. 401–404.

[24] E. Principi, S. Cifani, C. Rocchi, S. Squartini, and F. Piazza, "Keyword spotting based system for conversation fostering in tabletop scenarios: Preliminary evaluation," in *Proc. of HSI*, Catania, Italy, May 21-23 2009, pp. 216–219.

[25] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cavendish Laboratory, University of Cambridge, Tech. Rep., 2006.

[26] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[27] ITU-T, "Objective Measurement of Active Speech Level," ITU-T Recommendation P.56, Mar. 1993.

[28] H. Nanjo, H. Mikami, H. Kawano, and T. Nishiura, "A fundamental study of shouted speech for acoustic-based security system," in *Proc. of Interspeech*, Brighton, U.K., Sep. 6-10 2009, pp. 1027–1030.

[29] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, no. 54, pp. 732–742, 2012.

[30] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Proc. of ICASSP*, Prague, Czech Republic, May 22-27 2011, pp. 4968–4971.

[31] N. Obin, "Cries and whispers-classification of vocal effort in expressive speech," in *Proc. of Interspeech*, Portland, OR, USA, Sep. 9-13 2012, pp. 2234–2237.