Abstract
This demo demonstrates a small and portable system embedding microphone array processing and robust speech recognition for distant-speech interaction to control a floor lamp. The system is entirely contained inside the lamp and operates in real-time, “always-listening” mode. It runs on a small, low-power, fanless board and acts as the light control interface. The prototype shows the feasibility, potential and limits of the integration of speech technology in devices of everyday use.

Index Terms: microphone array, distant-talking speech recognition, embedded system.

1. Introduction
Controlling a device with spoken commands is a natural evolution of speech recognition technology applications. For specific users, it may represent a comfortable interaction mechanism that replaces the traditional remote control device. However, distant speech recognition [1] is still a challenging task, in particular when environmental noise and acoustics are hostile. Moreover, it may be characterized by a considerable computational complexity, due to the processing techniques necessary to obtain satisfactory performances in a wide range of conditions.

During the last decade, the progress of FBK technology for distant-speech recognition led to the development of systems that were deployed in different application contexts, mainly based on running algorithms on stand-alone PC platforms. Current efforts are also directed towards embedding this technology into ARM-based, low-cost, low-power consumption, miniaturized hardware platforms.

The proposed prototype is conceived to show this technological solution fully disappearing, integrated in a floor lamp, and invisible to the user, which can interact by voice at distance of some meters from the floor lamp. The front-end consists of an array of 8 digital MEMS microphones and related processing, while the remainder of the system runs on a Single Board Computer. Different tasks have been implemented on this platform: keyword spotting, command recognition and speech transcription. The prototype refers to the first case, i.e., a keyword spotting task running in an “always-listening” mode, which means that no push-to-talk button is required.

2. System architecture
The system is based on different processing components, as depicted in Figure 1. A front-end processing includes a Time Delay Estimation and a related Delay-and-Sum Beamformer. The output of the latter is then processed by a Voice Activity Detector in order to capture speech chunks to feed the recognizer. Speech recognition is finally performed in a keyword-spotting mode, based on acoustic models trained to ensure robust performance in various conditions.
2.2. Software Modules

The 8 channels are sampled at 48 kHz, but the front-end processing runs at 16 kHz through a standard downsampling provided by alsadc drivers. The software components have all been developed by FBK and are:

- **Time Delay Estimation (TDE)** continuously deriving 7 inter-channel delays (taking as reference one of the central microphones) based on GCC-PHAT [2].
- **Delay and Sum Beamforming** applied to input signals, based on the above-mentioned delays, and updated only if the actual coherence exceeds a given threshold.
- **Voice Activity Detection (VAD)**, based on joint use of energy and pitch information [3].
- **HMM Automatic Speech Recognition** based on cross-word context-dependent GMM acoustic models (about 3500 Gaussians) properly trained on contaminated signals [4][5] simulating an indoor environment with a reverberation time of about 0.6 seconds.

3. The recognition task

The recognition task comprises a list of 14 commands that allow one to change the light color, switch on/off the lamp, and obtain other effects (e.g., blinking, light intensity variations, etc.). The current version works in Italian and English. The interaction is triggered by the detection of a command as if it were a keyword, in order to allow the "always-listening" mode: a similar approach is being developed in DIRHA [6], a EU project where voice-based interaction in domestic environments is investigated. Because of the limited computational power of an ARM Single Board Computer, a major difficulty lies in being able to provide a real-time operation in a wide range of acoustic situations. The multi-microphone front-end process (acquisition and downsampling, TDE, Beamforming and VAD) involves a load of about 50% of one core. The decoder needs 80-90% of the second core, but it is a process that works on demand, depending on the presence of speech chunks provided by VAD. The standard response time of the floor lamp is less than 1 second. Latency may slightly increase due to a decrease in the SNR or in general for a mismatch between acoustic models and real acoustic environment. The parameter controlling the beam-search allows one to vary the promptness of the system: of course, faster reaction time can be obtained at the expense of a loss of recognition performance. In case of complex acoustics situations more computing power is occasionally required.

4. Conclusions

This paper provides an overview of a voice-controlled floor lamp. All components of the system are embedded into the lamp shade, thanks to the features offered by the ARM architecture and the digital MEMS microphones. This approach explores the way for the integration of speech technology in everyday devices, as alternative to a local-server based solution as that explored in the DIRHA project [6], or to a wireless-based, distributed system running in the cloud.

5. Acknowledgements

The authors would like to thank STMicroelectronics for kindly providing the MEMS microphones and the Snowball board.

6. References