# 2 INTELLIGENT ACOUSTIC INTERFACES

## Contents

**2.1 What is an intelligent acoustic interface? . . . . . . . . . . 12**

2.1.1 Acoustic interfaces . . . . . . . . . . . . . . . . . . . 12

2.1.2 The "intelligence" in interfaces . . . . . . . . . . . . 13

2.1.3 Human-machine interaction by intelligent acoustic interfaces . . . . . . . . . . . . . . . . . . . . . . . 14

2.1.4 Applications using intelligent acoustic interfaces . . 14

**2.2 Science and technology of intelligent acoustic interfaces 15**

2.2.1 Historical background on speech communications . 15

2.2.2 Philosophical background on intelligent interfaces . 18

**2.3 Intelligent acoustic interfaces for immersive communications . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20**

## 2.1 WHAT IS AN INTELLIGENT ACOUSTIC INTERFACE?

In order to formulate a comprehensive definition for the term “intelligent acoustic interface” (IAI), it is necessary first to know what an acoustic interface is and what makes an acoustic interface.

### 2.1.1 Acoustic interfaces

An *acoustic interface* provides a means to exchange acoustic information between two or more entities through an acoustic signal processing. More exactly, an acoustic interface is the front-end of a processing system of audio and speech signals aiming at the extraction and the reproduction of acoustic information. An acoustic interface is generally composed of a microphone array and one or more loudspeakers, as depicted in Fig. 2.1.

To control noise, reverberation, and competing speech, microphone array systems are generally more powerful than a single microphone [45]. Based on how the microphones are arranged, these systems have two basic forms: organized and distributed arrays [66]. In an *organized array*, the sensors are arranged to form a particular geometry (such as a line, a circle, or a sphere) in which each sensor’s position with reference to a common point is known. These sensors spatially sample the sound field and are required to have the same sensitivity. In comparison, a *distributed array* consists of randomly placed

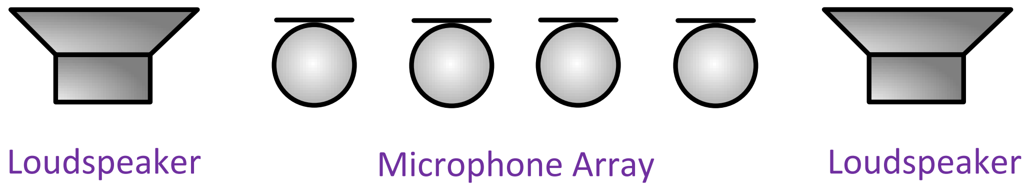

**Fig. 2.1:** *An acoustic interface.*

microphones. It offers the advantage of logistic convenience during installation and later operations. Typically, distributed arrays have a large number of elements forming a large sensor network. The microphone positions and the pattern of the array are usually not known, and a uniform response among the microphones cannot be presumed beforehand.

### 2.1.2 The "intelligence" in interfaces

"Intelligence" is not an easy term to define. What makes a system intelligent? In intelligent interfaces, the "intelligence" might be in predicting what the user wants to do, and presenting information with this prediction in mind [61]. Intelligent interfaces can also make doing a task more intuitive and helpful. Instead of trudging along a task in the mire of an inefficient and clumsy interface, the user might find a helpful and information-using interface to be more intelligent. Thus, "intelligence" does not actually mean cognition in this context; instead, it means using information in an appropriate manner [61, 87].

"Intelligence" in interfacing is a subjective term. One person may look at a system with context-sensitive help and say that the system seems smart; another person might look at the same system and see nothing special about it. In a sense, "intelligence" in interfaces might be defined as "the next best thing" [61]. Once we have a system which one would say is intelligent, the novelty of the system wears off, and people are in search for more intelligent interfaces. "Intelligence" is that goal which is always one step ahead of us; once we conquer it, it is no longer intelligence.

Interfaces can be intelligent about the user. Through the use of a user model, the system can tailor communication (both input and output) to the user [61]. Examples of tailored communications include methods of communicating (voice, visual, tactile) and way of presenting data (graph, chart, multimedia messages). The interface can also be sensitive to the wants and needs of the user. This ties closely with the user model, but it deals more with interface adaptability than outright use of models.

### 2.1.3 Human-machine interaction by intelligent acoustic interfaces

The *Association for Computing Machinery* defines *human-machine interaction* as “a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them” [62]. An important role in human-machine interaction is played by *intelligent acoustic interfaces* (IAIs). An IAI translates acoustic information from user to computer, and *vice versa*, in order to allow an homogeneous interaction between parties. From the user point of view, an IAI should be as invisible and intuitive as possible: working with and understanding an IAI should not be a task so that the user should be able to concentrate on the task which he is to perform.

An IAI must be able to adapt to user, to acquire and process information from user, to understand user requirements, to give user an answer satisfying his demands disguised as natural language or multimedia message. Moreover, once information has been acquired, an IAI must be able to autonomously decide whether the user needs an answer or not. In many cases, an IAI must learn user behaviour, mood and personality in order to yield an answer being as compliant as possible to user needs.

### 2.1.4 Applications using intelligent acoustic interfaces

IAIs are widely used in several fields of application, as also confirmed by the scientific and technological state of the art.

In the multimedia sector it is possible to think to applications such that: speech/audio real-time interaction [68]; speech automatic analysis, automatic music composition and transcription [15]; automatic genre and context recognition in broadcast programs [145]; high-interactivity entertainment [31], a sector that has viewed a growing interest also due to emerging videogame technologies.

In domotics IAIs may be employed in the following applications: the development of "intelligent rooms", in which speakers and speech commands must be recognized [142]; advanced anthropomorphic robotics [133]; integration with videosurveillance systems [33], that provide, in an automatic way, event identification, audio/video zoom in the region where the event is detected, and the consequent activation of an alarm or any other action related to the identified event.

Moreover, it is possible to exploits IAIs to develop aid systems for disabled people, that may be hearing aids or even devices able to provide an accurate reconstruction of an acoustic environment [122, 89].

## 2.2 SCIENCE AND TECHNOLOGY OF INTELLIGENT ACOUSTIC INTERFACES

### 2.2.1 Historical background on speech communications

Before the invention of electromagnetic telephones, there were mechanical devices for transmitting spoken words over a greater distance than that of normal speech. The very earliest mechanical telephones were based on sound transmission through *pipes* or other physical media (see Fig. 2.2). *Speaking tubes* long remained common, including a lengthy history of use aboard ships, and can still be found today.

The telephone emerged from the creation of, and successive improvements to the *electrical telegraph*. In 1804 Catalan polymath and scientist Francisco Salvá i Campillo constructed an electrochemical telegraph. An electromagnetic telegraph was created by Baron Schilling in 1832.

The first commercial electrical telegraph was constructed by Sir William Fothergill Cooke and entered use on the Great Western Railway in England. It ran for 13 miles from Paddington station to West Drayton and came into operation on April 9, 1839.

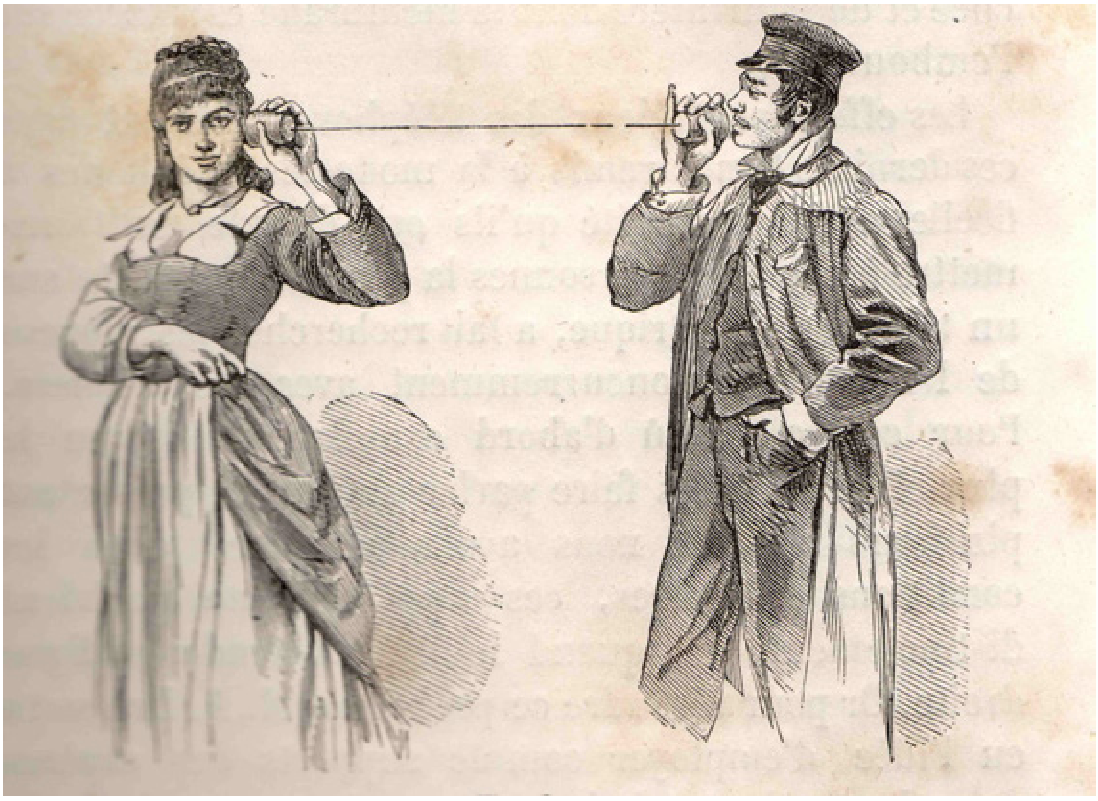

**Fig. 2.2:** *The tin can telephone, or also known as lover's phone, connected two diaphragms with a taut string or wire, which transmitted sound by mechanical vibrations from one to the other along the wire.*

During the second half of the 19th century inventors tried to find ways of sending multiple telegraph messages simultaneously over a single telegraph wire by using different modulated audio frequencies for each message. These inventors included Charles Bourseul, Thomas Edison, Elisha Gray, and Alexander Graham Bell. Their efforts to develop *acoustic telegraphy* in order to significantly reduce the cost of telegraph messages led directly to the invention of the *telephone*, or *the speaking telegraph*.

The commercial use of the telephone started in 1876 [57]. This was one year after Alexander Graham Bell filed a patent application for a telephone apparatus. Efforts to transmit voice by electric circuits, however, date further back. Already in 1854 Charles Bourseul described a transmission method. Antonio Meucci set up a telephone system in his home in 1855. Philipp Reis demonstrated his telephone in 1861. One has also to mention Elisha Gray who tragically filed his patent application for a telephone just 2 hours later than Alexander Graham Bell.

In the very early days of the telephone conducting a phone call meant to have both hands busy; one was occupied to hold the loudspeaker close to the ear and the other hand to position the microphone in front of the mouth. This troublesome way of operation was due to the lack of efficient electroacoustic converters and amplifiers. The inconvenience, however, guaranteed optimal conditions: a high signal-to-(environmental) noise ratio at the microphone input, a perfect coupling between loudspeaker and the ear of the listener, and - last but not least - a high attenuation between the loudspeaker and the microphone. The designers of modern speech communication systems still dream of getting back those conditions [57].

In a first step one hand had been freed by mounting the telephone device, including the microphone at a wall; further on, only one hand was busy holding the loudspeaker. In a next step the microphone and the loudspeaker were combined in a handset. Thus, still one hand was engaged. This basically remained the state of the art until today [57].

Early attempts to allow telephone calls with a loudspeaker and a microphone at some distance in front of the user had to use analog circuits. In 1957 Bell System introduced a so called *speakerphone*. At the same time, however, the telephone connection degraded to a half-duplex loop making natural conversations difficult. The introduction of a "center clipper" may be considered as a last step along this line [16]. This nonlinear device suppresses small amplitudes. Thus, it extinguishes small echoes. Moreover, small speech signals are erased, as well [57].

The invention of the least mean square algorithm in 1960 [153], the application of adaptive transversal filters [79, 130] and the availability of digital circuits with increasing processing power opened new paths to acoustic echo and noise control [57]. It took at least two more decades of breathtaking progress in digital technology until commercial applications of adaptive filters for acoustic echo and noise control became feasible.

Modern technologies are evolving towards new directions which take into account the *distant-talking*, i.e. the hands-free speech communication using intelligent acoustic interfaces. However, this change involves new challenging problems to address, as we see throughout this dissertation.

### 2.2.2 Philosophical background on intelligent interfaces

All along the human-interface interaction has aroused the interest of researchers, philosophers and cognitive scientists, which attempt to answer to questions about *artificial intelligence* (AI), such as "Can a machine display intelligence?".

The basic position of most AI researchers is summed up in this statement, which appeared in the proposal for the Dartmouth Conferences of 1956 [88]:

> *"Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."*

The first step to answering those questions is to clearly define "intelligence". Alan Turing, in a famous and seminal 1950 paper [144], reduced the problem of defining intelligence to a simple question about conversation. He suggests that:

> *"If a machine can answer any question put to it, using the same words that an ordinary person would, then we may call that machine intelligent."*

Recent AI research defines intelligence in terms of "intelligent agents", that is more close to our definition of "intelligent interfaces". An "agent" is something which perceives and acts in an environment; a "performance measure" defines what counts as success for the agent [117]:

> *"If an agent acts so as maximize the expected value of a performance measure based on past experience and knowledge then it is intelligent."*

In 1963, Allen Newell and Herbert Simon proposed that “symbol manipulation” was the essence of both human and machine intelligence. They wrote [95]:

> *“A physical symbol system has the necessary and sufficient means of general intelligent action.”*

This claim is very strong: it implies both that human thinking is a kind of symbol manipulation (because a symbol system is necessary for intelligence) and that machines can be intelligent (because a symbol system is sufficient for intelligence). Another version of this position was described by philosopher Hubert Dreyfus, who called it “the psychological assumption” [38]:

> *“The mind can be viewed as a device operating on bits of information according to formal rules.”*

A distinction is usually made between the kind of high level symbols that directly correspond with objects in the world and the more complex “symbols” that are present in a machine like a neural network. Moreover, Dreyfus argued that human intelligence and expertise depended primarily on unconscious instincts rather than conscious symbolic manipulation, and argued that these unconscious skills would never be captured in formal rules [38, 117].

Russell and Norvig point out [117] that, in the years since Dreyfus published his critique, progress has been made towards discovering the “rules” that govern unconscious reasoning. The situated movement in robotics research attempts to capture our unconscious skills at perception and attention [21]. Computational intelligence paradigms, such as *neural networks*, *evolutionary algorithms* and so on are mostly directed at simulated unconscious reasoning and learning.

Probably IAIs will never able to solve any problem that a person would solve by thinking, however, they may help users to enjoy an immersive communication.

## 2.3 INTELLIGENT ACOUSTIC INTERFACES FOR IMMERSIVE COMMUNICATIONS

After years of extraordinary technological advances in telecommunications, new requirements are demanded by users which are no longer satisfied with talking to someone over a long distance and in real time, but they want to collaborate through communication in a more productive way with the feeling of being together and sharing the same environment. That gives rise to *immersive communication*. Such immersive communication is yet to become a reality supported by modern communication technologies. A person's sense of acoustic immersion is formed by his or her sensory response to the auditory stimuli that exist in the ambiance of their environment [66].

Immersive communications take place in multisource environments, as depicted in Fig. 2.3 where interfering signals may degrade quality and intelligibility of the desired speech source. Therefore, acquisition of desired signals with high quality is far more difficult and challenging for immersive communications than in the classical telephony environment where the microphone is close to the user. In immersive communications, it is more likely that multiple parties will be involved and conferencing is a more common mode of operation than point-to-point calling. In conferencing, one may hear the unwanted interfering signals from every other participant and therefore the level of the perceived noise can grow with the number of participants. When the number is large and if interfering sources are not well controlled, the perceived noise can reach a level such that speech is overwhelmed. So interfering sources become a more quality-threatening problem for immersive voice communication [65, 66].

Immersive communication offers great opportunities for acoustic and speech signal processing and implies the use of IAIs. Voice is by far the dominant media in the exchange of conference content. In fact, a teleconferencing session can still go on when the video link is broken, but it has to stop if the

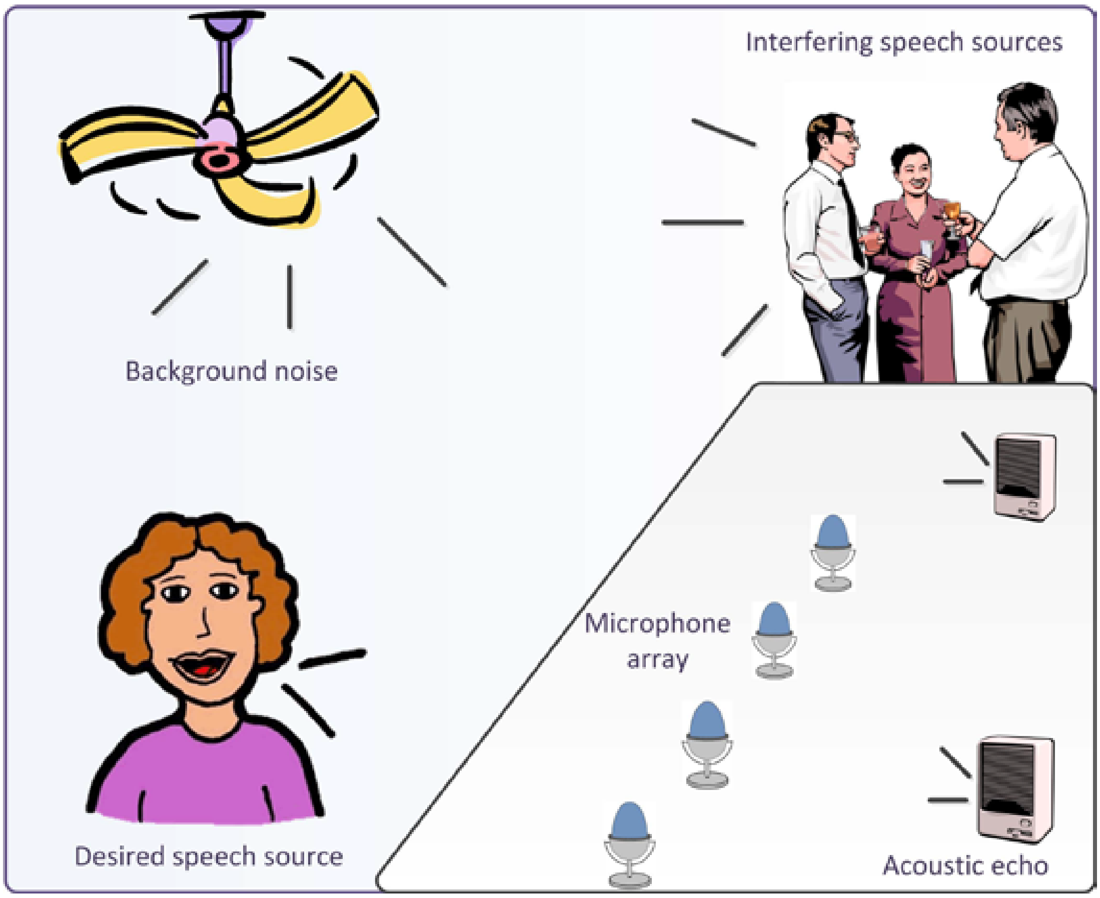

**Fig. 2.3:** *Immersive speech communication in multisource environment.*

audio link is disrupted. So in addition to the pursuit of multimodal capabilities, we should never forget the importance of *speech quality* (including intelligibility and naturalness) and *intermodal synergy*. Moreover, there are great potentials to improve these two factors in an immersive teleconference with multiple parties being involved since binaural hearing is now allowed and can be fully exploited. This is an imperative step towards immersive communication. With both ears being kept busy, our auditory system can more easily extract a single talker's speech among multiple conversations and background noise, and can more seamlessly work together with the visual system in an adverse acoustic environment for speech perception (e.g., lip-reading).

An IAI for immersive communication aims at extracting, from audio signals, useful informations for computational or human purpose, such as analysis or synthesis of audio signals. This feature is also known as *machine listening*. At the same time, an IAI has to reproduce desired acoustic information taking into account that the listener would hear the sound exactly as in the original sound field. This feature indeed is known as *spatial sound reproduction*. To these ends, an IAI needs to replicate four attributes of face-to-face communication [65, 66]:

1. full-duplex exchange;
2. freedom of movement without body-worn or tethered microphones (i.e., hands-free in the broad sense);
3. high-quality speech signals captured from a distance;
4. spatial realism of sound rendering.

These requirements imply that multiple microphones and loudspeakers would be used and the entire voice communication infrastructure might need to be renovated. However, the scope of this thesis mainly concerns with the machine listening feature, since we deal with adaptive algorithms which have to process the acoustic signals acquired by a microphone interface.