# A Biological Inspired Robotic Auditory System Based on Binaural Perception and Motor Theory

**Enzo Mumolo** and **Massimiliano Nolich**
DEEI, University of Trieste, Via Valerio 10, 34127 Trieste Italy


## Abstract

In this paper we present a novel artificial auditory system for humanoid robots. We address the problem of estimating an articulatory representation of the speech of the talker who is speaking to the robot using our auditory system. According to the motor theory of perception, the articulatory representation is the first step of a robust speech understanding process.

The system is composed by two parts, namely a beam-forming module and a perception module. The beam-former is two-channel (i.e. dual-microphones) and it is based on the super-directive beam-forming algorithm. The environment is scanned for seeking a sound source; when the direction of the source is found, the reception lobe of the dual-microphone system is steered to that direction and the signal is acquired. The perception module is based on a fuzzy computational model of human vocalization. In summary, the relationships between places of articulation and speech acoustic parameters are represented with fuzzy rules. Starting from the articulatory features, a set of acoustic parameters are generated according to the fuzzy rules. These acoustic parameters are used to generate a synthetic utterance which is compared in the perceptual domain to the corresponding spoken utterance. The goal of that is to estimate the membership degrees of the articulatory features using analysis-by-synthesis and genetic optimization.


## Introduction

It is well known that through the auditory system, living ceatures gather important information about the world in which they live. For lower animals, it may mean to be able to escape from a danger or to catch a prey, for humans it may mean to be able to focus one's attention on events, such as phone ringing, person talking etc. Robots also greatly benefit from auditory capabilities because their intelligence can be improved by fusing auditory information with the information coming from other sensors such as vision. The aim of this paper is to propose an artificial auditory system that gives a robot the ability to locate sounds sources using binaural perception, and to perceive speech, in terms of articulatory representation, on the basis of the motor theory of perception (Liberman & Mattingly 1985). In Fig. 1 the block diagram of our auditory system is reported. In summary, this paper focuses on the following two auditory capabilities

- binaural localization of human talkers
- speech perception by articulatory features estimation

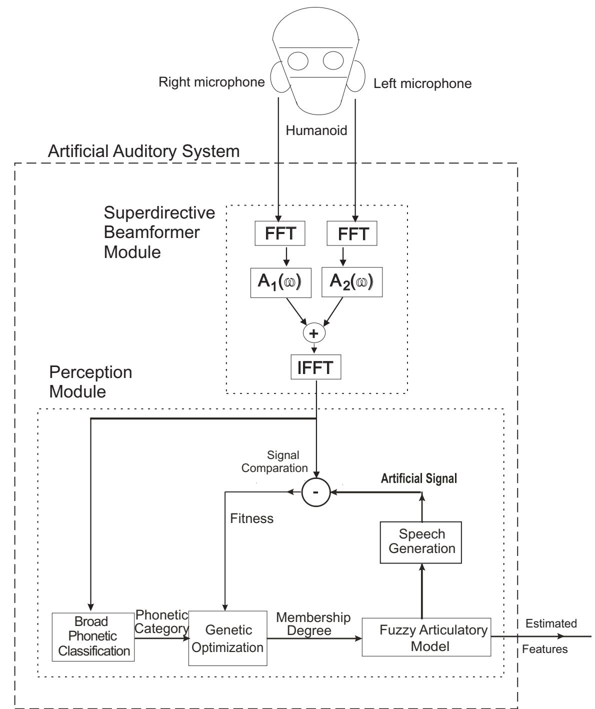


Figure 1: Block diagram of the artificial auditory system.

The environmental noise is reduced through a super-directive beam-former which is steered to the direction of the sound source. The obtained signal is then given as input to the perception module. We instrumented a mobile robot (Mumolo, Nolich, & Vercelli 2003) with a couple of microphones, at a distance of 15 cm, as reported in Fig. 2 where only the upper part of the system is shown. In this part the video and acoustic sensors are located.

Hearing is concerned with the sound processing mechanisms that occurs in the brain, and involves a very wide

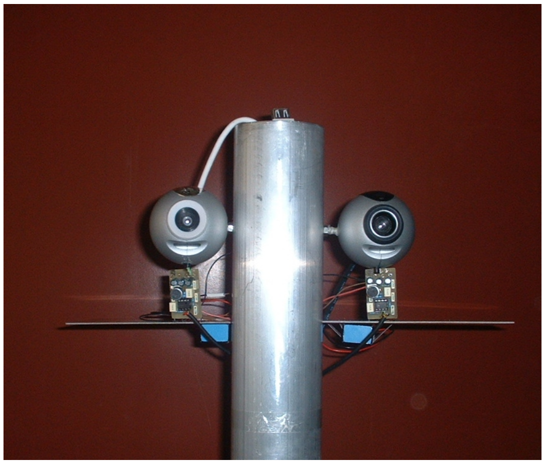

Figure 2: Picture of the prototype used in this work.

range of biological structures, from the transducers, which generally performs a kind of spectral representation of the input signal, to the brain where the high level functions are realized. Generally speaking, auditory capabilities include the approximate localization and tracking of sound sources, sound perception and classification, language recognition and the capacity to concentrate the attention on one flow of information, which allows human listeners to separate simultaneous talkers.

Clearly the human brain and even the brain of lower animals is by far too complex to mimic in an artificial auditory system. However, in this paper we study how and to what extent some speech perception theories can be included in an artificial auditory system.

It is worth briefly reviewing now some concepts of the speech perception theory which has, to some extent, inspired this work. The Motor Theory of Speech Perception, though severely criticized, is one of the most widely-cited theories of human speech perception. According to Liberman & Mattingly this theory states that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators. According to this theory, when we perceive speech, we perceive the gestures that correspond to the articulatory movements of the speaker, such as lip rounding and jaw raising. Furthermore, in this theory there is a *specialized module* in the brain that translates from the acoustic signal to the intended articulatory gestures. According to Bell *et al.*, such a module might work using the analysis-by-synthesis method (Bell *et al.* 1961), in which a mental model of a speech synthesizer is used to generate various acoustic properties. The acoustic-gesture parameters that are input to this synthesizer are varied until the error between the synthesized acoustic properties and the observed acoustic properties is minimized.

Our perception module is based on fuzzy logic. The possibility of facing the vagueness involved in the interpretation of phonetic features using methods based on fuzzy logic has been realized in the past, when approaches to speech recognition via phonetic classification were proposed (De Mori 1983; De Mori & Laface 1980).

In this work, we developed fuzzy rules which relate places of articulation with the corresponding acoustic parameters. The system requires a training phase for estimating the degrees of membership. Thus, the algorithm tries to reproduce some input speech and, in this way, the articulatory characteristics of the speaker who trained the system are learned.

## Previous work

Artificial auditory systems in robotics is a quite recent field of research, and in the following we report a few significant results.

At Kyoto University the SIG robots have been developed since 2000. These robots have increasing binaural auditory capacity. Latest developments, besides localization, have been directed to sound separation and speech recognition of the separated signals (Yamamoto *et al.* 2004). Other systems use more than two microphones. The robot described in (Choi *et al.* 2003) uses eight microphones organized in a circular array and performs beamforming and speech recognition.

## Talker localization

The estimation of the direction of arrival of a talker is performed by scanning the environment and computing the probability of vocalization on the signal coming from that direction. Assuming that the noise is Gaussian, (Soon, Koh, & Yeo 1999) report that the probability of vocalization can be computed as:

$$P(H_1|V) = e^{-\xi_k} I_0(2\sqrt{\gamma_k \xi_k}/(1 + e^{-\xi_k} I_0(2\sqrt{\gamma_k \xi_k}) \quad (1)$$

where $H_1$ is the speech presence hypothesis, V is the measured signal envelope, $\xi_k$ is the a-priori SNR, $\gamma_k$ is the a-posteriori SNR, and $I_0$ is the zero-order modified Bessel function.

### The super-directive beam-forming algorithm

A general description of a two channel beamformer is given by: $y(n) = \sum_{i=0}^{1} x_i(n) * a_i$, where $x_i(n)$ is the input signal and * denotes convolution with the filter $a_i$, which realizes beam-forming. In the frequency domain, beam-forming becomes $Y(\omega) = \sum_{i=0}^{1} X_i(\omega) A_i(\omega)$. In the super-directive beamformer described in (Bitzer, Simmer, & Kammeyer 1999), the filter A is given by:

$$A = \frac{\Gamma^{-1} d}{d \Gamma^{-1} d} \quad (2)$$

where $\Gamma$ is a coherence matrix whose elements are the normalized cross-power spectral density of the two sensors and $d$ is the steering vector. In Fig. 3 we report the diagram of the beam received at the beamformer of our auditory system.

## The perception module

### Preliminaries

According to the distinctive feature theory (Fant 1973), phonemes are classified in terms of manner and place of articulation. The manner of articulation is concerned with the

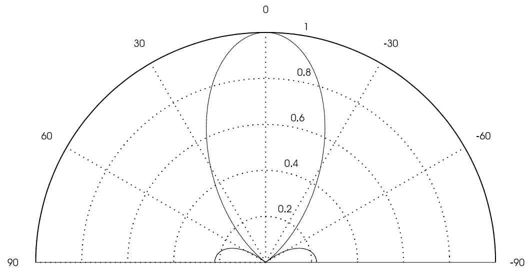


Figure 3: Beam pattern at 1546 Hz of the super-directive beamformer.

degree of constriction imposed by the vocal tract on the airflow. On the other hand, place of articulation refers to the location of the most narrow constriction in the vocal tract. Using manner and place of articulation, any phoneme can be fully characterized in binary form. The idea of our work, instead, is that the features are not binary, but fuzzy: as a matter of fact, a certain degree of fuzzyness, due to the lack of knowledge, is involved in their characterization, which thus should be fuzzy rather than strictly binary. For example, it may be that the /b/ phoneme, classically described as plosive, bilabial and voiced in binary forms, may involve also a certain degree of anteriority and rounding, as well as some other features. In fact, our approach can be viewed as an extension of classical distinctive feature theory where all sounds of human languages can be described by a set of binary features where binary means that the feature in question is present or not.

The following phonemes in Italian language were considered in this work: vowel, liquid, nasal, fricative, plosive, affricate. Such phonemes are described using the following six manners of articulation: vowel, nasal, fricative, plosive, affricate, liquid.

Moreover, the following places of articulation for the vowels have been used: open, anterior, sonorant and round. A vowel opening is related to the distance of the tongue from the palate, which is as far as possible; the anterior sounds are produced with the tongue at or in front of the alveolar ridge. A sonorant sound is produced when the vocal cords vibrate, while the round vowels are produced with a considerable degree of lip rounding.

Similarly to the vowels, the places of articulation for the consonants are related to the place where the vocal tract is narrower. We consider the following places of articulation: bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental and velar.

The next important issue related to this work is the relation between articulatory and acoustic features. The acoustic features, such as the formant frequencies, are obviously related to the dimension and constrictions of the oral and laryngeal phonetic organs (Ladefoged 1996; Pickett 1999). For example, raising the tongue to the palate enlarges the pharyngeal cavity and decreases the volume of the oral cavity. As a result, F1 is low and F2 is high. In the same way, retracting and lowering the tongue lengthens the oral cavity and decreases the pharyngeal cavity. As a result, F1 is high and F2 is low and so on.

It is worth noting that, in this work, we have described this knowledge into fuzzy rules (in a way described for example as: ”for each first phoneme, if the second phoneme is open, then F1 is medium”, or ”for each first phoneme, if the second phoneme is anterior, then F2 is medium-high”. The rules require some basic phonetic knowledge which is language dependant, but they don’t need to be more detailed than those described. What makes things work, is the subsequent optimization which find the best values of the membership degrees which are used in the actual synthesis.

## The perception algorithm

The perception module is described in Fig. 1. According to the block diagram, we now summarize the actions of the algorithm.

The system requires a supervised learning phase which works as follows: the operator pronounces a word and the robot generates an artificial replica of the word based on the articulatory and acoustic estimation. This process iterates until the artificial word matches the original one according to the operator’s judgement. At this point the speech learning process is completed and the robot has learnt how to pronounce that word in terms of articulatory movements. It is worth noting that several repetitions of the same phoneme in different contexts are needed to improve performances.

The synthesis of the output speech is performed using a reduced Klatt formant synthesizer (Klatt 1987).

Since the fuzzy rules, however, describe the locus of the acoustical parameters, a model of the parameters profiles has been introduced. The profile of each synthesis parameter ‘p’ is described with four control features, namely the initial and final intervals I(p) and F(p), the duration D(p) and the locus L(p). The I(p) control feature determines the length of the starting section of the transition, whose slope and target values are given by the D(p) and L(p) features. The parameter holds the value specified by their locus for an interval equal to F(p) ms; however, if other parameters have not completed their dynamic, the final interval F(p) is prolonged. The I(p), F(p), and D(p) parameters are expressed in milliseconds, while the target depends on what synthesis control parameter is involved; for example, for frequencies and bandwidths the locus is expressed in Hz, while for amplitudes in dB.

## Phoneme and Control Parameters Fuzzification

As mentioned above, the phonemes are classified into broad classes by means of the manner of articulation; then, the place of articulation is estimated by genetic optimization. Therefore, each phoneme is described by an array of nineteen articulatory features, six of them are boolean variables and represent the manner of articulation and the remaining thirteen are fuzzy and represent the place of articulation. In this way, our approach appears as an extension of the classical binary definition; in our case a certain vagueness in the definition of the places of articulation of the phonemes is introduced.

Representing the array of features as (vowel, plosive, fricative, affricate, liquid, nasal | any, rounded, open, anterior, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar). The /a/ phoneme, for example, can be represented by the array:

$$[1, 0, 0, 0, 0, 0|1, 0.32, 0.9, 0.12, 1, 0, 0, 0, 0, 0, 0, 0, 0]$$

indicating that /a/ is a vowel, with a degree of opening of $0.9$, of rounding of $0.32$, and it is anterior at a $0.12$ degree. The /b/ phoneme, on the other hand, can be considered a plosive sonorant phoneme, bilabial and slightly velar, and therefore it can be represented by the following array:

$$[0, 1, 0, 0, 0, 0|1, 0, 0, 0, 0.8, 0.9, 0, 0, 0, 0, 0, 0, 0.2].$$

The arrays reported as an example have been partitioned for indicating the boolean and the fuzzy fields respectively. Such arrays, defined for each phoneme, are the membership values of the fuzzy places of articulation of the phonemes.

The output of the phonetic module, described in the following, is given in terms of these four parameters; hence, the translation to the synthesis parameters trajectories required by the synthesizer must be performed.

On the other hand, the I, D, F and L fuzzy variables, defined in a continuous universe of discourse, can take any value in their interval of definition. The fuzzy sets for these variables have been defined as follows:

- Duration D(p). The global range of this fuzzy variable is 0-130 ms, with trapezoidal membership functions. Such values are indicated as follows:

$$\text{Very Short, Medium Short, Short, Medium,} \\ \text{Medium Long, Long, Very Long} \quad (3)$$

- Initial Interval I(p). As D(p), this fuzzy variable is divided into trapezoidal membership functions in a 0-130 ms interval. The fuzzy values are indicated, in this case:

$$\text{Instantaneous, Immediate, Quick, Medium,} \\ \text{Medium Delayed, Delayed, Very Much Delayed} \quad (4)$$

- Final Interval F(p). The numeric range is 0–130 ms and the fuzzy values are the same as indicated for the Initial Interval I(p).
- Locus L(p). The fuzzy values of this variable depend on the actual parameter to be controlled. For AV, AH and AF the fuzzy values are:

$$\text{Zero, Very Low, Low, Medium Low, Medium,} \\ \text{Medium High, High, Very High} \quad (5)$$

and their membership functions are equally distributed between 12 and 80 dB with the trapezoidal shape. The other gain factors, namely A2F-A6F and AB, take one of the following values:

$$\text{Very Low, Low, Medium Low, Medium,} \\ \text{Medium High, High, Very High} \quad (6)$$

in the range 0-80 dB with the same trapezoidal shape as before. The values of L(F1), L(F2) and L(F3) are named as in (6), with trapezoidal membership functions uniformly distributed from 180 to 1300 Hz, 550 to 3000 Hz and 1200 to 4800 Hz for the first, second and third formants respectively. Finally, the loci of the bandwidths B1, B2 and B3 take one of the fuzzy values described in (6), and their trapezoidal membership functions are regularly distributed in the intervals 30-1000 Hz for B1, 40-1000 Hz for B2 and 60-1000 Hz for B3.

**Fuzzy Rules and Defuzzification** By using linguistic expressions which combine the above linguistic variables with fuzzy operators, it is possible to formalize the relationship between articulatory and acoustic features.

In general, the rules involve the actual and the future phonemes. Moreover, the fuzzy expressions involve the fuzzy operators AND, NOT and OR. Since the manner of articulation well partitions the phonemes in separated regions, the rules have been organized in banks, one for each manner.

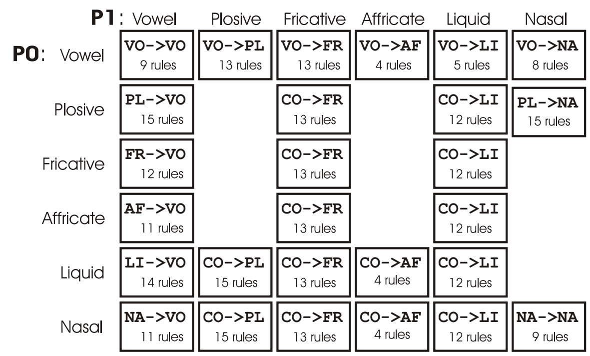


Figure 4: Outline of the bank of fuzzy rules. P0 and P1 represent the actual and target phonetic categories. CO denotes a generic consonant.

That is, calling P0 and P1 the actual and the future phonemes respectively, the set of rules is summarized in Fig. 4. The rule decoding process is completed by the defuzzification operation, which is performed with the fuzzy centroid approach.

As an example, the fuzzy rules related to F1 and F2 for the silence-nasal transition are the following.

```
IF ( P0 IS ANY AND P1 IS ANY ) THEN
  { L(F1) IS VERY_LOW ; }
IF ( P0 IS ANY AND P1 IS PALATAL ) THEN
  { L(F2) IS MEDIUM;
    L(F2) IS MEDIUM_HIGH; }
IF ( P0 IS ANY AND P1 IS DENTAL ) THEN
  { L(F2) IS LOW ;
  L(F2) IS MEDIUM_LOW; }
IF ( P0 IS NOT SON AND P1 IS DENTAL ) THEN
  { L(F2) IS MEDIUM; }
IF ( P0 IS ANY AND P1 IS BILABIAL ) THEN
  { L(F2) IS VERY_LOW; }
```

Concluding, as shown in Fig. 4, there are several transitions which are performed with the same set of rules. For example, all the transition toward fricatives and liquid phonemes are realized with the same bank of rules. This is because the related transitions can be approximated with a strong discontinuity, and thus they can be considered independent from the starting phonemes; the symbol ‘CO’ used in these banks stands, in fact, for a generic consonant sounds. Other banks are missing; this is because they are concerned with transitions which occur very rarely in Italian language.

**Genetic optimization of articulatory and acoustic parameters** Let us take a look again at Fig. 1. Genetic optimization estimates the optimum values of the degrees of membership for the articulatory features used to generate an artificial replica of the input signal by comparing the artificial with the real signal.

**Genetic optimization module** The optimal membership degrees of the articulatory places minimize the distance from the uttered signal; the inputs are the number of phonemes of the signal and their classification in terms of manner of articulation.

One of the most important issues of the genetic algorithm is coding. The chromosome used for the genetic optimization of a sequence of three phonemes is shown in Fig. 5. It represents the binary coding of the degrees of membership. The genetic algorithm uses only mutations of the chromosome. This means that each bit of the chromosome is changed at random and the mutation rate is constant to 2%.

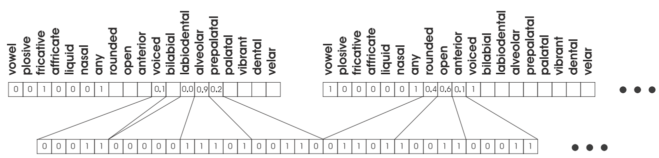


Figure 5: The binary chromosome obtained by coding.

**Fitness computation and articulatory constraints** An important aspect of this algorithm is the fitness computation, which is represented by the big circle symbol in Fig. 1. The fitness, which is the distance measure between original and artificial utterances and is optimized by the genetic algorithm, is an objective measure that reflects the subjective quality of the artificially generated signal. For this purpose we used the Modified Bark Spectral Distortion (MBSD) measure (Wang, Sekey, & Gersho 1992; Yang, Dixon, & Yantorno 1997). This measure is based on the computation of the pitch loudness, which is a psychoacoustical term defined as the magnitude of the auditory sensation. In addition to this, a noise masking threshold estimation is considered. This measure is used to compare the artificial signal generated by the fuzzy module and the speech generation module against the original input signal.

The MBSD measure is frame based. That is, the original and the artificial utterances are first aligned and then divided into frames and the average squared Euclidean distance between spectral vectors obtained via critical band filters is computed. The alignment between the original and artificial utterances is performed by using dynamic programming (Sakoe & Chiba 1978), with slope weighting as described in (Rabiner & Juang 1993).

Therefore, using the mapping curve between the two signals obtained with dynamic programming, the MBSD distance $D$ between original and artificial utterances represented respectively with $X$ and $Y$ is computed as follows:

$$D(X,Y) = \frac{1}{L_\Phi} \sum_{k=0}^{T}$$

$$\left[ \sum_{j=0}^{K} M(\Phi_y(k), j) \left| L_x(\Phi_x(k), j) - L_y(\Phi_y(k), j) \right| m(k) \right]$$

where $T$ is the number of frames, $K$ is the number of critical bands, $\Phi = (\Phi_x, \Phi_y)$ is the non-linear mapping obtained with dynamic programming, $L_\Phi$ is the length of the map, $L_x(i,j)$ is the Bark spectrum of the $i$-th frame of the original utterance, $L_y(i,j)$ is the Bark spectrum of the $i$-th frame of the artificial utterance, $M(i,j)$ is the indicator of perceptible distortion at the $i$-th frame and $j$-th critical band, and $m(k)$ are the weights. The coefficient $M(i,j)$ is a noise masking threshold estimation which determines if the distortion is perceptible by comparing the loudness of the original and artificial utterances.

The fitness function of the Place of Articulation (PA), i.e. the measure to be maximized by the genetic algorithm, is then computed as:

$$Fitness\,(PA) = \frac{1}{D(X,Y)} + \sum_{j=1}^{N_c} P_j.$$

where $P_j$ is the $j$-th penalty function and $N_c$ is the number of constraints. In fact, in order to correctly solve the inverse articulatory problem, several constraints, due to the physiology of the articulations, have been added to the fitness.

In conclusion, the optimization of places of articulation (PA) can be expressed as follows:

$$PA = argmax\,\{Fitness\,(PA)\}\,.$$

## Experimental results

The membership degrees of the phonemes are estimated from speech and included in the matrix reported in Fig. 6.

| | Round | Open | Anterior | Sonorant | Bilaial | Labiodental | Alveolar | Prepalatal | Palatal | Vibrant | Dental | Velar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| /sil/ | | | | | | | | | | | | |
| /a/ | | | | | | | | | | | | |
| ⋮ | | | | | | ⋮ | | | | | | |
| /n/ | 0 | 0 | 0 | 0.8 | 0.6 | 0 | 0 | 0 | 0.8 | 0 | 0.4 | 0 |
| /o/ | 0.6 | 0.8 | 0.5 | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ⋮ | | | | | | ⋮ | | | | | | |
| /z/ | | | | | | | | | | | | |

Figure 6: Membership degrees of phoneme transitions coming from 'any' phoneme. The membership degrees for the utterance 'no' are shown.

In Fig. 7 the dynamic of the estimated membership degrees of the articulatory places of articulation for the Italian word 'nove' is reported.

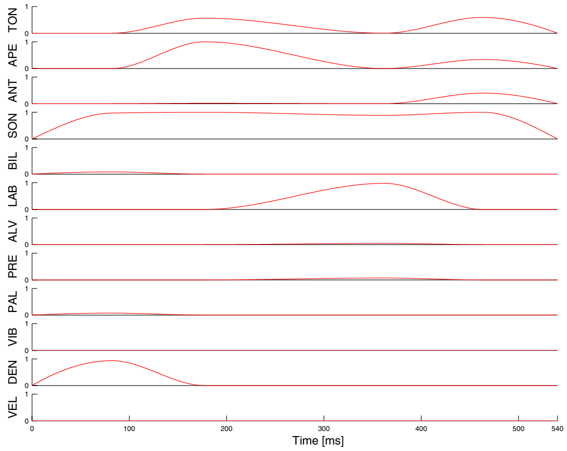


Figure 7: Places of articulation of the Italian word 'nove' estimated with the perception module.

## Final remarks and conclusions

In this paper we have dealt with auditory systems for humanoid robots. Our approach is based on a beam-forming module and a perception module. The beam-former is based on the super-directive algorithm of (Bitzer, Simmer, & Kammeyer 1999) while the perception module is a new approach which uses a set of fuzzy rules and a genetic algorithm for the optimization of the degrees of membership of the places of articulation. The membership values of the places of articulation of the spoken phonemes have been computed by means of genetic optimization.

The proposed auditory system has many potential applications in robotics: first of all the super-directive beam-former can identify the direction of arrival of a sound source - in order to facilitate approaching manoeuvre of the mobile robot - and it allows the acquisition of a noise-suppressed signal, which is important in distant talking context. Then, the perception module estimates acoustic and articulatory features on the basis of the motor theory of perception.

Aa additional outcome of our auditory system is the production of artificial speech which mimics the input signal. It is worth noting that, besides their use is speech recognition, the estimated articulatory features can be used for controlling the mechanical parts of talking heads in humanoid robotics.

At the present state of this research, much care must be put in the module 'Broad Phonetic Classification' which estimates the phonetic categories from speech because there isn't yet a mechanism which correct false classifications.

## References


Bell, C.; Fujisaki, H.; Heinz, J.; Stevens, K.; and House, A. 1961. Reduction of Speech by Analysis-by-Synthesis Techniques. *Journal of the Acoustical Society of America* 33:pp.1725–1736.

Bitzer, J.; Simmer, K.; and Kammeyer, K. 1999. Multi-Microphone Noise Reduction Techniques for Hands-Free Speech Recognition - A Comparative Study. In *Proc. Robust Methods for Speech Recognition in Adverse Conditions (ROBUST-99)*, pp.171–174.

Choi, C.; Kong, D.; Kim, J.; and Bang, S. 2003. Speech enhancement and recognition using circular microphone array for service robots. In *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.3516–3521.

De Mori, R., and Laface, P. 1980. Use of Fuzzy Algorithms for Phonetic and Phonemic Labeling of Continuous Speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol.2:pp.136–148.

De Mori, R. 1983. *Computer Models of Speech Using Fuzzy Algorithms*. New York: Plenum Publishing Corporation.

Fant, G. 1973. *Speech Sounds and Features*. MIT Press.

Klatt, D. 1987. Review of Text-to-Speech Conversion in English. *J. Acous. Soc. Am.* pp.737–793.

Ladefoged, P. 1996. *Elements of Acoustic Phonetics.* Chicago: University of Chicago Press.

Liberman, A., and Mattingly, I. 1985. The Motor Theory of Speech Perception Revised. *Cognition* 21:pp.1–36.

Mumolo, E.; Nolich, M.; and Vercelli, G. 2003. Algorithms for acoustic localization based on microphone array in service robotics. *Robotics and Autonomous Systems* Vol.42(2):pp.69–88.

Pickett, J. 1999. *The Acoustic of Speech Communication: Fundamentals, Speech Perception Theory, and Technology.* Boston: Allyn and Bacon.

Rabiner, L., and Juang, B. 1993. *Fundamentals of Speech Recognition*. Prentice–Hall.

Sakoe, H., and Chiba, S. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust. Speech Signal Processing* Vol.26:pp.43–49.

Soon, I.; Koh, S. N.; and Yeo, C. K. 1999. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. *Signal Processing* 75:pp.151–159.

Wang, S.; Sekey, A.; and Gersho, A. 1992. An Objective Measure for Predicting Subjective Quality of Speech Coders. *IEEE J. on Select. Areas in Comm.* Vol.10.

Yamamoto, S.; Nakadai, K.; Tsujino, H.; and Okuno, H. 2004. Improvement of robot audition by interfacing sound source separation and automatic speech recognition and missing feature theory. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.2111–2116.

Yang, W.; Dixon, M.; and Yantorno, R. 1997. A modified bark spectral distortion measure which uses noise masking threshold. In *IEEE Workshop on Speech Coding For Telecommunications Proceeding*, pp.55–56.