Speaker Distance Detection Using a Single Microphone

Georganti, Eleftheria; May, Tobias; van de Par, Steven; Harma, Aki; Mourjopoulos, John

Published in:
I E E E Transactions on Audio, Speech and Language Processing

Link to article, DOI:
10.1109/TASL.2011.2104953

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Georganti, E., May, T., van de Par, S., Harma, A., & Mourjopoulos, J. (2011). Speaker Distance Detection Using a Single Microphone. *I E E E Transactions on Audio, Speech and Language Processing*, *19*(7), 1949-1961. https://doi.org/10.1109/TASL.2011.2104953

Speaker Distance Detection Using a Single Microphone

Eleftheria Georganti, Tobias May, Steven van de Par, Aki Härmä, and John Mourjopoulos, *Member, IEEE*

***Abstract*—A method to detect the distance of a speaker from a single microphone in a room environment is proposed. Several features, related to statistical parameters of speech source excitation signals, are introduced and are shown to depend on the distance between source and receiver. Those features are used to train a pattern recognizer for distance detection. The method is tested using a database of speech recordings in four rooms with different acoustical properties. Performance is shown to be independent of the signal gain and level, but depends on the reverberation time and the characteristics of the room. Overall, the system performs well especially for close distances and for rooms with low reverberation time and it appears to be robust to small distance mismatches. Finally, a listening test is conducted in order to compare the results of the proposed method to the performance of human listeners.**

***Index Terms*—Acoustic signal processing, distance measurement, room acoustics.**

I. Introduction

METHODS for speaker localization and distance detection have a broad range of applications, such as intelligent hearing aid devices [1], speech recognition [2], auditory scene analysis [3], [4], augmented reality audio [5] and hands-free communication systems [6]–[8]. In this paper, we focus on the applications of distributed hands-free (or *ambient*) telephone systems [9]. Ambient telephones consist of a central unit and arrays of small hands-free terminal devices distributed in a multi-room environment (see [8] for the details). Processing and rendering speech signals captured by such array allows having hands-free phone calls while the user is moving from one room to another [10] and multiple simultaneous calls can be placed in different parts of the environment. In this respect, the ambient telephone aims at simulating the real physical presence of a remote caller in the user's environment. The ambient telephone system can be controlled automatically if the users and their active conversations are tracked in the environment [11], achieved via combinations of different techniques using microphones, cameras, and other sensors. In this paper, we develop a method for detecting the distance of the local user from an ambient telephone terminal unit based on the received single microphone signal. Knowing the distance between terminals and user would allow to select the terminal which is closest to the user and presumably has the best signal-to-noise ratio.

A common approach to such source localization and distance detection tasks, is the use of a microphone array and to perform time delay estimation (TDE), using, e.g., the generalized cross-correlation (GCC) algorithm [12]. The angle of arrival can be calculated from the TDE and applying the triangulation rule can lead to the bearing estimation. This basic bearing estimation process forms the foundation of most of the source-localization techniques, even though many algorithms may formulate and solve the problem from a different theoretical perspective [13]. Lately, research work on the localization problem has been undertaken using binaural signals [14], [15]. These methods utilize auditory cues that underlie distance judgments by humans. Such listeners' abilities to determine source distance under reverberant conditions have been extensively studied [16]–[26] and they have initiated novel techniques for the localization problem and especially for distance estimation using only two sensors [14], [15], [27]–[29].

However, an ambient telephone terminal device should be a small and low-power device with limited computational resources and for this reason it is preferable if all localization processing is performed in the central unit, which then only receives monophonic microphone signals or the output of a fixed beamformer. In a calibrated ambient telephone system, user localization is always possible to some extent based on the TDEs between the microphone signals. However, in such a scenario, the positions and orientations of all terminals should be known and even then, there are often detection ambiguities due to geometric constraints and the small number of devices in individual rooms. Hence, the information about the distance of a talker from each terminal would be very valuable in resolving such ambiguities and especially in cases when the user is, for the most of the time, significantly closer to one device than the other devices.

Recently, there has been some work on the estimation of the talker/microphone distance using binaural signals. Lu *et al.* [15], [29] have proposed a binaural distance estimator for the case where the receiver is moving. Smaragdis and Boufounos

Manuscript received December 16, 2009; revised June 19, 2010; accepted December 22, 2010. Date of publication January 10, 2011; date of current version July 15, 2011. This work was supported by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Stategic Reference Framework (ESPA) 2007-2013, according to Contract MICRO2-38. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian Mak.

E. Georganti is with the Audio and Acoustic Technology Group, Wire Communications Laboratory, Electrical and Computer Engineering Department, University of Patras, 26500 Rio, Greece (e-mail: egeorganti@upatras.gr).

T. May and Dr. S. van de Par are with the Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: tobias.may@uni-oldenburg.de; steven.van.de.par@uni-oldenburg.de).

Dr. A. Härmä is with the Digital Signal Processing Group, Philips Research Europe, Eindhoven, NL-5656 AE, The Netherlands (e-mail: aki.harma@philips.com).

J. Mourjopoulos is with the Audio and Acoustic Technology Group, Wire Communications Laboratory, Electrical and Computer Engineering Department, University of Patras, 26500 Rio, Greece (e-mail: mourjop@upatras.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2104953

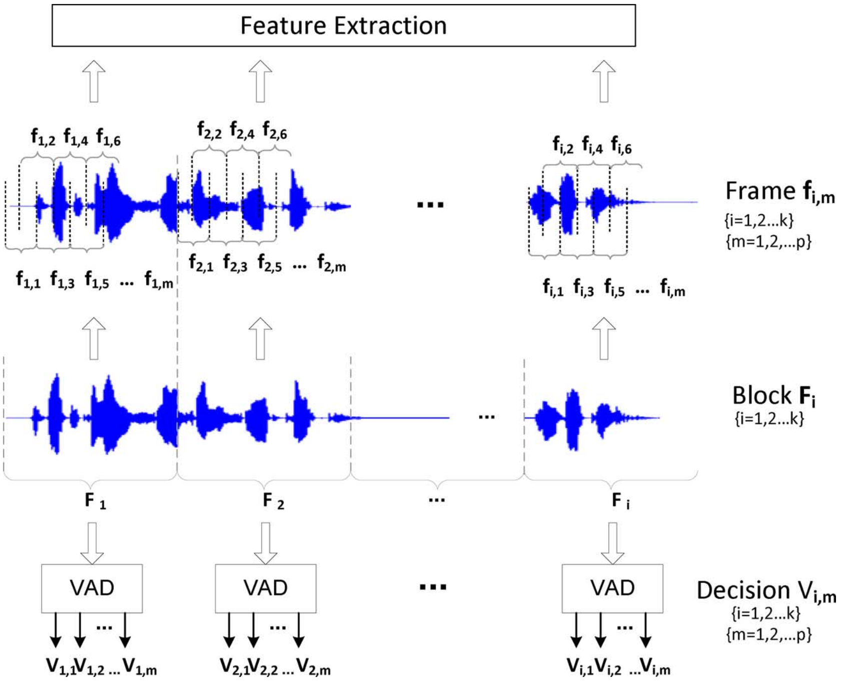


Fig. 1. Input speech segmentation. The speech signal is processed in blocks ($F_1, F_2, \ldots F_i$). Each block F_i has a duration of 2 s. In order to extract the features of each block, it is divided into frames $f_{i,m}$ of 20 ms using 50% overlap. Speech blocks (F_i) are passed through a voice activity detector (VAD) [33] that assigns a value ($V_{i,m}$) equal to 1 if the frame contains speech and a 0 value if not.

[28] have employed an expectation maximization algorithm that learns the amplitude and phase differences of cross-spectra in order to recognize the position of a sound source using two microphones. This method was later improved by Vesa [14] in order to account for the positions that have the same azimuth angle. Lately, there has been some work using monophonic signals, such as the work of Lesser and Ellis [30], where hand claps are classified as near-field or far-field based on a few simple features such as the center of mass, the slope of decay and the energy compared to background noise, but this method is applicable for transients only. Other existing monophonic techniques [31], [32] mainly focus on the estimation of direction (angle detection) between the source (talker) and the receiver (microphone) and to the best of the authors knowledge, there has not been any work in the past for talker/microphone absolute distance detection from received monophonic speech signals.

The present method is based on previous and recent findings related to the effects of the reverberant energy on the statistics of signals. It is well known that the source/receiver distance affects significantly the signal properties, being largely manifested as variation of the direct-to-reverberant ratio. Statistics of the spectral magnitude of anechoic and reverberant signals (speech/audio) and some of the effects of reverberant energy on the statistics of speech as a function of distance have been studied in [34]–[36]. Recently, several speech and audio dereverberation techniques rely on such statistical findings in order to extract the interfering noise-reverberation distortion from the audio-speech signal [37]–[44].

In this paper, the distance-dependent variation of several temporal and spectral statistical features of single-channel signals is studied. A novel sound source distance detector, based on these features is developed and its performance is evaluated in different acoustic environments.

This paper is organized as follows. In Section II, the proposed method for distance detection is described and the features are defined and analyzed. Section III gives the description of the classifier using Gaussian mixture models (GMMs). The experimental evaluation of the method is given in Section IV and the proposed method is compared to two other methods in Section V. Finally, the performance of the method is then compared to the performance of human listeners (Section VI) and the paper concludes with a summary of the present work.

II. Distance Features Extraction

The received speech signals, sampled at 44.1 kHz, are segmented in blocks (F_i) of 2 s, from which 20-ms frames ($f_{i,m}$) are extracted, using 50% overlap (see Fig. 1). Additionally, speech signals are passed through a voice activity detector (VAD) [33] that returns the VAD decision sample-by-sample. The VAD decision is then segmented in the same way as the speech signals and a value ($V_{i,m}$) equal to 1 (detected speech activity) is assigned if 60% of the VAD decision samples of one frame contains speech and 0 otherwise. After the speech signal is segmented, the frames are processed by the feature extraction scheme. The block diagram of the feature extraction is shown in Fig. 2 and consists of two processing blocks (Block I and II). In Block I, after the speech segmentation described by Fig. 1, if the processed frame ($f_{i,m}$) contains speech (the assigned value $V_{i,m}$ from the VAD is equal to 1), the speech signal is Hanning-windowed and this frame is used for the feature extraction being further processed in Block II. Otherwise the frame data are ignored. Signal features are extracted only from the frames (called "subfeatures") that contain speech. Then, the histogram of these features over each 2-s block is computed. For each block one set of features is then calculated.

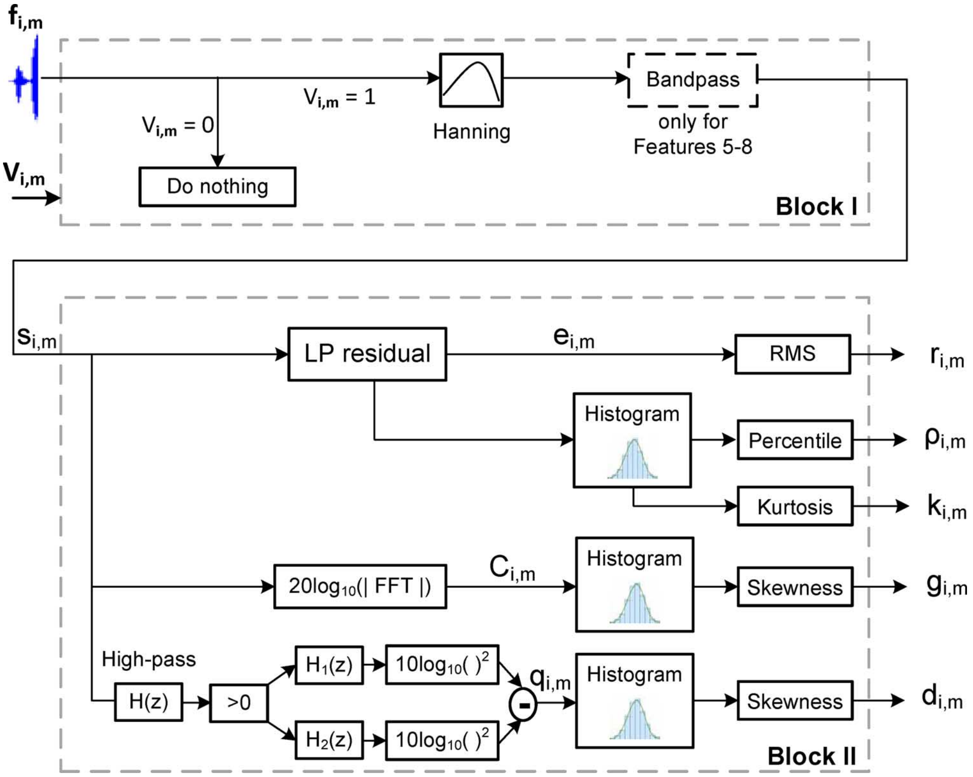


Fig. 2. Processing scheme for the subfeatures extraction. In Block I, after the speech segmentation described by Fig. 1, if the processed frame ($f_{i,m}$) contains speech, the speech signal is Hanning-windowed and this frame is used for the feature extraction being further processed in Block II.

In Sections II-A–II-D, the processing of Block II is described analytically for the feature extraction.

A. *Linear Prediction Residual Peaks*

Linear prediction (LP) analysis [45] of order p is carried out on the speech samples $s_p(n)$ of each frame $f_{i,m}$. Let $\hat{s}_p(n)$ denote the predicted signal, a_k the kth LP coefficient and $A(z)$ the LP inverse filter. The speech signal filtered with this inverse filter gives the LP residual signal $e_{i,m}(n)$.

As is well known, clean voiced speech is typically modeled using a source-filter model, where the glottal excitation signal $G(z)$ is filtered by the acoustic transfer function of the vocal tract $T(z)$. The propagation of the sound from the speaker's lips to the microphone is modeled by the acoustic transfer function $H(z)$. Therefore, the received microphone signal can be modeled as $S(z) = H(z)T(z)G(z)$.

Moreover, by definition, the all-pole LPC model, given by $1/A(z)$, models efficiently any minimum-phase system such as the vocal tract transfer function $T(z)$. Therefore, the residual signal $E(z)$ can be approximated by

$$\begin{aligned} E(z) = A(z)S(z) &= A(z)H(z)T(z)G(z) \\ &\approx A(z)H(z)\frac{1}{A(z)}G(z) \\ &\approx H(z)G(z) \end{aligned} \tag{1}$$

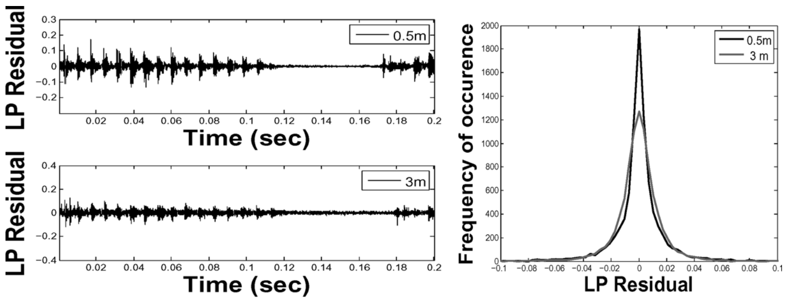
so that the influence of the vocal tract is largely eliminated from the output. However, $H(z)$ is poorly modeled with the linear prediction, because of the length of the room impulse response, $h(n)$. Since the glottal excitation signal consists of a sequence of brief wide band pulses in the time domain, the effect of the early part of the room impulse response should be clearly "visible" in the residual signal in between the maximum peaks. This implies, that for clean voiced speech, the LP residual signal displays strong peaks corresponding to glottal pulses, whereas for reverberated speech such peaks are more spread in time due to room reflections.

Fig. 3(a) shows the tenth-order LP residual for a block of a speech signal (2 s) recorded at 0 m (top) and 3 m (bottom). The two signals are normalized to have root mean square (rms) value equal to 1. In Fig. 3(b), the histograms of the LP residual amplitude values of those two signals can be seen. It can be noted that the amplitude values of the signal recorded at 3 m are more spread in time and the corresponding histogram is less peaked compared to the 0 m histogram. This effect is independent of the signal gain as the two signals are normalized to have rms equal to 1.

Thus, a measure of the amplitude spread of the LP residual can serve as metric registering the amount of reverberation in the signal [37] and consequently as distance metric, since distance affects the direct-to-reverberant ratio (DRR) [46]. Clearly, distance also affects the signal gain and as will be shown below, this aspect is addressed by appropriate preprocessing.

Based on the above findings, a feature related to the LP residual is calculated here. After the processing of Block I in Fig. 2, if the assigned value $V_{i,m}$ (from the VAD) of the frame is 1, the rms

$$r_{i,m} = \sqrt{\frac{1}{\tau}\sum_{n=1}^{\tau} e_{i,m}^2(n)} \tag{2}$$



(a) (b)

Fig. 3. (a) Tenth-order LP residual for a block of speech signal recorded at 0.5 m (top) and 3 m (bottom) from the microphone and (b) the corresponding histograms of the amplitude values of the LP residual of the speech recordings at the distances of 0.5 m and 3 m.

of the LP residual (for $p = 10$-th order), $e_{i,m}(n)$, of the frame is calculated, where τ is the number of samples of the frame.

A signal value below which 90% of the observations may be found is determined (percentile value of 0.9). In this paper this measure is called the percentile value $\rho_{i,m}$. The subfeature values of $\rho_{i,m}$ and $r_{i,m}$ are calculated for the frames $f_{i,m}$ that contain speech (the assigned value $V_{i,m}$ of the frame $f_{i,m}$ from the VAD should be equal to 1) of block F_i.

Then, the subfeatures percentile $\rho_{i,m}$ and rms $r_{i,m}$ are summed over the frames

$$P_{F_i} = \sum_{m=1}^{p} \rho_{i,m} \quad \text{and} \quad R_{F_i} = \sum_{m=1}^{p} r_{i,m}. \tag{3}$$

Finally, the feature of the LP residual peak for the 2 s block F_i is defined as the ratio of the subfeatures percentile $p_{i,m}$ and rms $r_{i,m}$, i.e.,

$$\textbf{Feature 1}: LPratio(F_i) = \frac{P_{F_i}}{R_{F_i}}. \tag{4}$$

In this way, the extraction of the feature, is made independent of the signal gain, but as it will be shown, depends on the distance between source and receiver.

B. Linear Prediction Residual Kurtosis

The second feature is also based on the characteristics of the LP residual using a similar approach as the one described in Section II-A. In Fig. 3(b), it can be seen that the histogram of the LP residual values at 3 m is less peaked compared to the 0-m histogram. This property can be utilized using the statistical quantity of kurtosis, which is a measure of whether the data are peaked or flat relative to a normal distribution. For this feature, the kurtosis $k_{i,m}$ of the LP residual amplitude values $e_{i,m}(n)$, [38], [41], of each frame $f_{i,m}$, is computed according to

$$k_{i,m} = \frac{\tau \sum_{n=1}^{\tau} (e_{i,m}(n) - \bar{e}_{i,m})^4}{\left(\sum_{n=1}^{\tau} (e_{i,m}(n) - \bar{e}_{i,m})^2\right)^2} \tag{5}$$

where $\bar{e}_{i,m}$ indicates the sample average of the LP residual $e_{i,m}(n)$ (see Fig. 2). Then, in order to calculate the feature for the whole block F_i, the mean of the subfeature values of the kurtosis $k_{i,m}$ is taken:

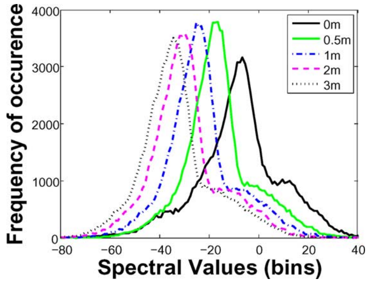


Fig. 4. Spectrum magnitude statistics of a reverberant signal recorded at different distances from the source for a 2-s block.

$$\textbf{Feature 2}: Kurt(F_i) = \frac{1}{p} \sum_{m=1}^{p} k_{i,m}. \tag{6}$$

C. Skewness of the Spectrum

The third feature explores the effect of reverberation on the spectral characteristics of speech [34]–[36]. In Fig. 4, the spectrum magnitude statistics of a reverberant signal (duration of 2 s), recorded at four different distances in a typical room is shown. The recorded signals were normalized to a maximum value of 0 dB full scale (FS). It can be seen that as the distance increases, the histograms of the spectral values are more biased to lower spectral values and present longer right tail values. This observed asymmetry of the histograms can be quantified by using the statistical quantity of skewness. For the third proposed feature, the power spectrum of the speech frame is expressed in dB, $C_{i,m}$ and the subfeature skewness $g_{i,m}$ is computed (see Fig. 2), according to

$$g_{i,m} = \frac{\sqrt{\tau} \sum_{n=1}^{\tau} (C_{i,m}(z) - \bar{C}_{i,m})^3}{\left(\sum_{n=1}^{\tau} (C_{i,m}(z) - \bar{C}_{i,m})^2\right)^{3/2}} \tag{7}$$

where $\bar{C}_{i,m}$ indicates the sample average of the power spectrum $C_{i,m}(z)$, of the frame. Then, in order to calculate the feature of the whole block F_i, the mean of the values of skewness is taken as

$$\textbf{Feature 3}: SpecSkew(F_i) = \frac{1}{p}\sum_{m=1}^{p} g_{i,m}. \tag{8}$$

D. Skewness of Energy Differences

This feature is mainly based on an onset detector and on empirical considerations. First, in order to remove the dc component, the signal is passed through a high-pass filter $H(z)$ with the following transfer function

$$H(z) = \frac{1 - z^{-1}}{1 + 0.9z^{-1}} \tag{9}$$

in order to remove the dc component. Then, half-wave rectification is performed and the signal is filtered with the filters $H_1(z)$ and $H_2(z)$, respectively (see Fig. 2), where

$$H_1(z) = \frac{1}{1 - a_1 z^{-1}},\ a_1 = 0.99 \tag{10}$$

$$H_2(z) = \frac{1}{(1 - a_2 z^{-1})^3},\ a_2 = 0.998. \tag{11}$$

The length of the impulse response of $H_1(z)$ is significantly shorter than that of $H_2(z)$ and focused on the most recent sample values. Therefore, the ratio between the outputs of $H_1(z)$ and $H_2(z)$ can be used as a detector for sharp changes in the signal amplitude and as demonstrated in Fig. 2:

$$q_{i,m}(n) = s_{i,m}(n) * h(n) * \left[10 \cdot \log_{10} \frac{h_1^2(n)}{h_2^2(n)}\right] \tag{12}$$

where $h(n)$, $h_1(n)$, and $h_2(n)$ are the impulse responses of the corresponding filters [see (9)–(11)] and $s_{i,m}(n)$ is the input frame. In the case of reverberation, one may assume that signals will have sharper onsets than offsets, because offsets are blurred by the tail of the room impulse response. Therefore, it can be assumed that the shape of the sample value distribution of $q_{i,m}(n)$ will depend on the amount of reverberation in the signal, independently of the actual input signal. The property of the distribution used in the current paper is the skewness of the filtered energy differences $d_{i,m}$ (see Fig. 2), which is calculated as

$$d_{i,m} = \frac{\sqrt{\tau}\sum_{n=1}^{\tau}(q_{i,m}(n) - \bar{q}_{i,m})^3}{\left(\sum_{n=1}^{\tau}(q_{i,m}(n) - \bar{q}_{i,m})^2\right)^{3/2}}. \tag{13}$$

In order to calculate the feature of the whole block F_i, the mean of the skewness values is taken over the 2-s block

$$\textbf{Feature 4}: FiltSkew(F_i) = \frac{1}{p}\sum_{m=1}^{p} d_{i,m}. \tag{14}$$

E. Band-Pass Filtered Features

The four features that were described in the previous sections were calculated using the full frequency range of the signal. Additionally, the same features were extracted for a high-frequency bandpass filtered version of the signals. Following the procedure shown in Fig. 2, after the Hanning-window, a bandpass filter (with cutoff frequencies: 10 kHz and 15 kHz) is applied and four extra features are extracted (see Fig. 2). The empirical consideration of choosing the high-frequency cutoff frequency of the bandpass filter was that at far distances, air absorption decreases more the level of high frequencies compared to the lower frequencies [16]. This specific bandpass filter was chosen after several tests with other bandpass filters (5–10 kHz, 15–20 kHz), because it gave the highest performance gain compared to other cutoff frequencies. From now on, these bandlimited features will be referred to as follows:
Feature 5: LPratioBP(F_i), corresponding to feature of (4).
Feature 6: KurtBP(F_i), corresponding to feature of (6).
Feature 7: SpecSkewBP(F_i), corresponding to feature of (8).
Feature 8: FiltSkewBP(F_i), corresponding to feature of (14).
The ending "BP" is used to denote their bandpass characteristic.

III. Distance Model

Gaussian mixture models (GMMs) can be used to approximate arbitrarily complex distributions and are therefore chosen to model the distance-depending distribution of the extracted features [47], [48].

A. Model Initialization

In this paper, five different classes (λ) corresponding to the test distances were chosen (0 m, 0.5 m, 1 m, 2 m, 3 m) and the eight different features described in Section II, are used. Each class is represented by a GMM and is referred to with its three parameter sets (μ_i, Σ_i, p_i).

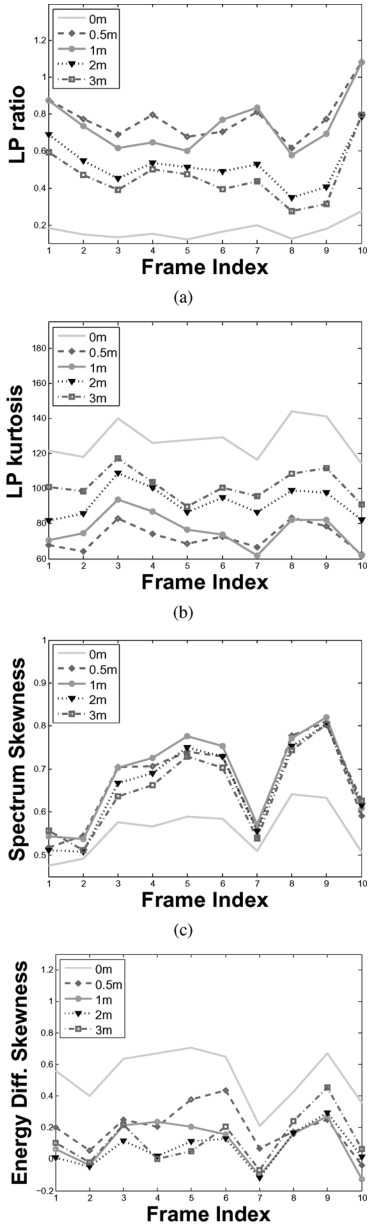
The GMM was initialized using the k-means algorithm [49]. The expectation–maximization (EM) algorithm [47] was used to estimate the set of GMM parameters with a maximum number of 300 iterations. The system was trained using the eight features, described in Section II. Twenty Gaussian components were used, diagonal covariance matrices and the complete feature space was normalized across all classes to have zero mean and unit variance. The normalization was done before training the classifier and the corresponding normalization values were stored for each feature dimension independently. After the training phase, the trained GMMs were rescaled to fit the original range of the feature space as it was before normalization. In this way, the GMM training is not biased due to the different range of the feature dimensions and no normalization is required in the testing stage. The distance detection, derived by the system, was then evaluated for the four rooms listed in Table I using speaker-dependent and speaker-independent distance models.

B. Feature Space

Fig. 5 shows typical extracted values of the four features described in Section II for a specific speech signal recorded at different distances from the microphone (0 m, 0.5 m, 1 m, 2 m, 3 m). Fig. 5(a) and (b) shows clearly the dependence of the feature's values on the distance. In contrast, the feature values of Fig. 5(c) and (d) do not follow such a clear trend, but it was

TABLE I
GEOMETRICAL AND ACOUSTICAL CHARACTERISTICS OF THE ROOMS

Room	Volume (m^3)	RT (s)	d_{crit} (m)	Description
A	60	0.39	0.7	Small office
B	555	0.84	1.5	Classroom
C	1292	1	2	Small auditorium
D	9633	1.47	4.5	Large auditorium



(a)

(b)

(c)

(d)

Fig. 5. Typical extracted features as a function of the block under examination. Each block contains 2 s of speech. (a) LP residual peaks ratio. (b) LP residual kurtosis. (c) Skewness of the spectrum. (d) Skewness of energy differences.

found that they still have an added value, when combined with the rest of the features in the pattern recognizer.

In Fig. 6, the histograms of the extracted values of the LPratio feature for Rooms A and D (see Table I) can be seen. The feature value for Room A [Fig. 6(a)] indicates clear dependency on the distance. On the other hand, the same feature for Room D [Fig. 6(b)] overlaps for all the distance classes, apart from the 0-m class.

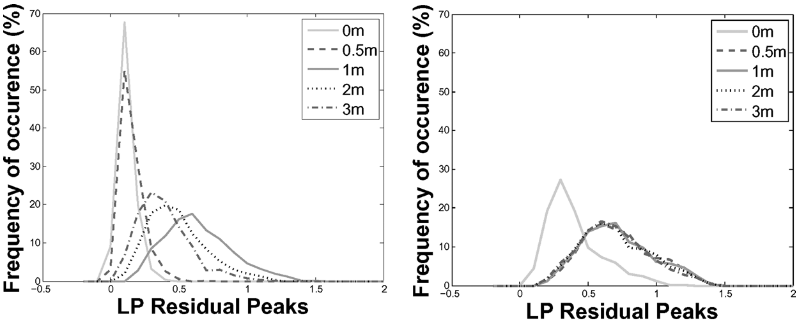
IV. EXPERIMENTAL EVALUATION

A. Database

In order to train and evaluate the system, several speech recordings were taken in an anechoic chamber located at Philips Research Laboratories, Eindhoven. For the recordings, 16 speakers (4 female and 12 male) had to read a piece of text for 3 minutes and their speech was captured at the distance of 0.5 m using an omnidirectional measurement microphone at the sampling frequency of 44.1 kHz. This sampling frequency was chosen after several tests with different sampling frequencies (16 kHz, 22.05 kHz, 44.1 kHz) which show that 44.1 kHz led to the highest performance of the method. Half of the recordings (24 minutes) were used for the training the other half for the testing stage (24 minutes). Then, these dry recordings were convolved with impulse responses (IR) that were measured at different distances (0 m, 0.5 m, 1 m, 2 m, 3 m) between source and receiver in four different rooms, hence simulating the presence of the speakers at those positions within the room. The range of these distances was chosen bearing in mind the potential application of the method on an ambient telephone system, where the distance between the possible placement of devices and seating positions is usually less than 3 m. The resolution of 1 m would be enough for such a system, where each terminal makes an independent estimate of the distance to the speaker, in order to resolve ambiguities due to the geometric constraints of the room. However, after the closest terminal is determined, a higher resolution (0.5 m) at close distances (less than 1 m) would be also useful for capturing and reproducing the signal. Note that the 0-m IR measurement was actually taken at a distance of 5 cm, but for reasons of clarity this class is denoted as the 0-m class. The volume, reverberation time (RT) and the critical distance (d_{crit}) of the rooms can be seen in Table I. For the case of Room A, two extra sets of recordings were taken. First, the receivers were placed at 1.5 m, 2.5 m, and 3.5 m from the source and secondly they were offset (by 10 cm) (see Fig. 9) with respect to the initial positions. The purpose of these measurements was to evaluate the system in conditions with small or significant placement mismatches compared to the positions of the training stage.

B. Feature Selection

In order to examine the effectiveness and the importance of the described features, the performance of the method was initially evaluated for each feature individually and the results can be found in Table II. The method was initially tested in Room A, which is a typical room in a home environment. It can be noted that the most effective features are the LPratio and the KurtBP. Furthermore, combinations of the four features LPratio, Kurt, SpecSkew, and FiltSkew were evaluated using the full frequency range and the bandpass-limited frequency range. This combination of features led to an increase of the performance and especially when using the full band version of the features (69.7%). Maximum performance (75.4%) was achieved, when all the eight features were employed, thus in the proposed method all of them are used. Here, it should be stated that the above tests were performed using noise-free data



(a) (b)

Fig. 6. Histograms of the LPratio feature values for Room A ($\mathrm{RT} = 0.39$ s) and Room D ($\mathrm{RT} = 1.47$ s). (a) Room A. (b) Room D.

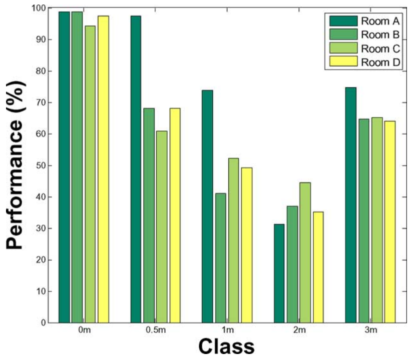


Fig. 7. Performance of the method as a function of distance for the four different rooms using speaker-independent speech model.

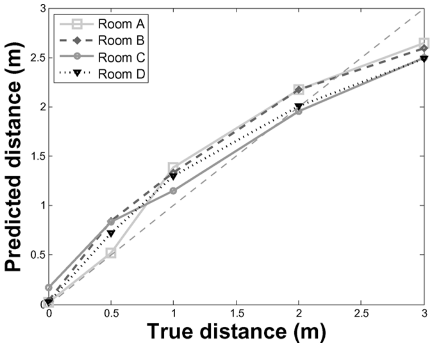


Fig. 8. Predicted distance as a function of true distance using speaker-independent speech model. The predicted distance is calculated from the confusion matrices, according to (15).

and that in practice, the bandpass features are effective only if the signal-to-noise-ratio in this frequency range is sufficient.

C. Block Duration Selection

The performance of the method was tested using blocks of 2 s, 4 s, and 8 s duration. Increasing the block size led to higher performance in the case of Rooms B, C, and D, but not for the case of Room A. This indicates that a block duration of 2 s is sufficient for a room with short RT, such as Room A, where the acoustics influence less the speech signals. As the proposed method is mainly based on the statistical properties of speech, longer blocks are expected to lead to more robust calculation of the feature values, but this may lead to increased latency for any real-time implementation. In the context of ambient telephony, the system is expected to respond sufficiently fast, e.g., when the user is moving in the room, thus for the evaluation of the proposed method the block of 2-s duration was chosen.

TABLE II
TYPICAL PERFORMANCE OF THE METHOD USING INDIVIDUAL AND VARIOUS COMBINATIONS OF FEATURES. THE HIGHEST PERFORMANCE WAS ACHIEVED USING ALL 8 FEATURES

Features	Performance
LPratio	49 %
Kurt	44.9 %
SpecSkew	38 %
FiltSkew	28.4 %
LPratioBP	41.4 %
KurtBP	46.3 %
SpecSkewBP	39 %
FiltSkewBP	31.8 %
LPratio, Kurt, SpecSkew, FiltSkew	69.7 %
LPratioBP, KurtBP, SpecSkewBP, FiltSkewBP	64.4 %
All 8 features	75.4 %

D. Speaker-Independent Performance

The system was trained separately for each room using half of the recordings, which were randomly chosen. Thus, the recordings of six male and two female speakers were used to extract the features for the training stage. The features were extracted for the five different classes (corresponding to distances 0 m, 0.5 m, 1 m, 2 m, 3 m) using the procedure described in Section II.

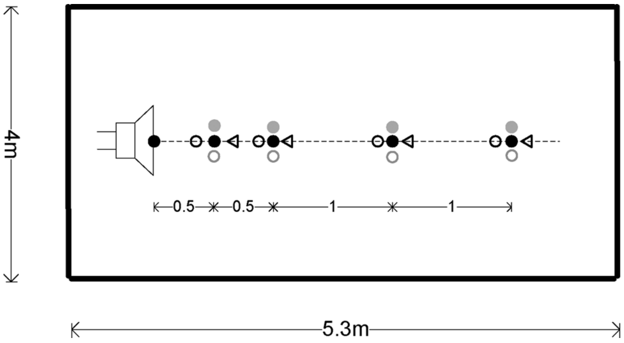
For the evaluation of the system the remaining half of the recordings was used. The recordings for the training and the evaluation stage were randomly selected, assuring that the chosen speakers for the training stage were always different to the ones chosen for the evaluation stage. The features were extracted following the procedure, described in Section II and Table III(a)–(d) show the performance of the method for the four rooms as confusion matrices.

Moreover, Fig. 7 presents the performance of the method per class, which can be derived by plotting the diagonal of the confusion matrices.

Fig. 8 shows the predicted distance as a function of the true distance. The predicted distance Δ_j for the jth class, is calculated from the confusion matrices of Table III, according to

$$\Delta_j = \sum_{j=1}^{5} \epsilon_{i,j} \lambda_i \tag{15}$$

where $\epsilon_{i,j}$ is the i, jth element of the confusion matrix, i represents the actual class, j the predicted and λ_j is the distance corresponding to the jth class (0 m, 0.5 m, 1 m, 2 m, 3 m). It can be seen that performance depends both on the distance and on the acoustical properties of the rooms. The method performs



	Symbol	Distances (m)
Train	●	0, 0.5, 1, 2, 3 (On Axis)
Test A	○	0, 0.4, 0.9, 1.9, 2.9
Test B	◁	0, 0.6, 1.1, 2.1, 3.1
Test C	●	0, 0.5, 1, 2, 3 (10cm Right)
Test D	○	0, 0.5, 1, 2, 3 (10cm Left)

Fig. 9. Ground plan of Room A showing the positions of the measurements.

TABLE III
PERFORMANCE OF THE METHOD FOR (a) ROOM A, (b) ROOM B, (c) ROOM C, (d) ROOM D, USING CONFUSION MATRICES. THE ROWS REPRESENT THE ACTUAL CLASSES (a) AND THE COLUMNS THE PREDICTED CLASSES (p). (a) ROOM A. (b) ROOM B. (c) ROOM C. (d) ROOM D

(a)

A/P	0 m	0.5 m	1 m	2 m	3 m
0 m	99	1	0	0	0
0.5 m	2	97	0	0	1
1 m	0	0	74	14	12
2 m	0	0	25	32	43
3 m	2	1	6	16	75

(b)

A/P	0 m	0.5 m	1 m	2 m	3 m
0 m	99	0	0	0	1
0.5 m	0	68	20	5	7
1 m	0	26	41	19	14
2 m	0	5	16	37	42
3 m	0	0	4	31	65

(c)

A/P	0 m	0.5 m	1 m	2 m	3 m
0 m	94	0	0	1	5
0.5 m	0	61	28	8	3
1 m	0	27	52	14	7
2 m	1	10	16	45	28
3 m	2	4	6	23	65

(d)

A/P	0 m	0.5 m	1 m	2 m	3 m
0 m	98	2	0	0	0
0.5 m	2	68	25	5	0
1 m	2	17	49	24	8
2 m	2	3	24	36	35
3 m	3	1	7	25	64

better in Room A, where the reverberation time is the lowest ($\mathrm{RT} = 0.39$ s) and its robustness decreases for rooms with more reverberation. However, the method performs better in Room C than in Room B, even though the critical distance is farther and the RT longer. Metrics such as the RT of the room or the critical distance seem not to be sufficient to characterize their effect on the performance of the method, because they only contain information about the total absorption and the volume of the room, but not about the exact geometry of the room or the first reflections. As can be observed from Fig. 6(a) the LP ratio feature values for Room A clearly depend on the distance and they only slightly overlap. On the other hand, in Fig. 6(b) the same feature values for Room D appear to overlap for the classes beyond 1 m. Thus, it becomes difficult for the pattern recognizer to form a robust prediction model for those classes in this room.

This can be explained, considering the fact that the sound field (acoustic transfer function) at the microphone consists of a direct and a reverberant component. The level of the direct component depends on distance (decreasing with $1/r^2$), whereas the reverberant component does not depend on distance. For close distances, the signal properties will change with increasing distance, as discussed earlier. On the other hand, as distance increases, the sound field is dominated by the statistical reverberant field and hence very little changes will be detected on the signal properties with increasing distance and this could explain the fact that the features would start to overlap. Although one would expect that such an effect would depend on the critical distance and that the features would start to overlap beyond this distance the results here do not unequivocally indicate such dependency and further investigation is needed.

From the results of Table III, the 0-m class is classified with the highest performance, which is above 94% for all the rooms. However, the performance drops to below 40% for the 2-m class, even for the least reverberant room (Room A), since this class appears to be mostly confused with the 3-m class and to some lesser extend with the 1-m class.

According to the preceding discussion, either the distances of 2-m and 3-m approach or belong to the reverberant sound field of the room and this appears to result to an “overlapping” of the feature values at such distances as the distance increases.

E. Speaker-Dependent Performance

In the context of ambient telephony system, only a specific set of people may use the system, and therefore, it is interesting to see if the system performance improves for such a speaker-dependent case. For this reason, the method was also tested using the same speakers for the training and the evaluation stage, but using different speech material (phrases).

As can be seen in Table IV, the performance of the method slightly increases by up to 1.1% for Room A, having the highest increase. In such case, since in a small room the room acoustics influence the speech signals less, the method appears to be more sensitive to the individual speaker. In the case of the rooms (Room B, Room C, Room D) with longer reverberation time (see Table I), the performance of the method remains the same,

TABLE IV
COMPARISON OF THE PERFORMANCE OF THE METHOD USING SPEAKER-INDEPENDENT AND SPEAKER-SPECIFIC SPEECH MODEL

Room	Different Speaker	Same Speaker
A	75.4 %	76.5 %
B	62 %	62.5 %
C	63.4 %	63.4 %
D	63 %	63.4 %

TABLE V
MEAN PERFORMANCE OF THE PROPOSED METHOD, WHEN THE SYSTEM IS TRAINED IN ONE SPECIFIC ROOM AND TESTED IN ANOTHER ONE

Test	Training	Testing	Mean performance
Test I	Room D	Room A	29.6 %
Test II	Room A	Room B	39.7 %
Test III	Room C	Room B	58.1 %

as can be seen in Table IV. Here, the specific speaker signal appears to be less critical as the room acoustics have much stronger effect on the received signals.

F. Dependence on the Room and Position

In this section, the performance of the proposed method is tested for different rooms and positions than those used for training, since it is of interest to know whether the proposed approach is applicable to previously unseen situations.

1) Room Mismatch: For this experiment, the system classifier was trained in one room and then its performance was then tested in another room. In Table V, the results for three different tests are shown. For the first test (I), the system is trained in Room D (large auditorium, $\mathrm{RT} = 1.57$ s) and then tested in Room A (small office, $\mathrm{RT} = 0.39$ s). The performance of the method drops significantly (from 75.4% to 29.6%) and the method fails to classify correctly. Similar behavior of the method is observed for the second test (II), where the classifier for Room A is tested in Room B, although in this case the performance drops less (from 62% to 39.7%). On the other hand, when the system is trained in Room C and tested in Room B (Test III), the performance drops to 58.1% from the initial performance of 62%. Note that at the first test (I) the difference of the RT between the two rooms was 1.18 s and the performance reduction was 45.8%. At the second test (II), the RTs of the rooms differ by 0.45 s and the performance reduction was 22.3%. Finally, for the third test (III), where the difference of the RT of the rooms was much lower (0.16 s), the performance drop was also lower, i.e., only 3.9%.

The above results indicate that the method is sensitive to the RT of the room and it is essential to train the system in a room with similar acoustical properties to the room in which the system should be used. This is expected, since the features employed by the method are sensitive to the reverberant energy of the signals. Furthermore, it is clear that the time and spectral properties of the recorded speech signal at a distance of 1 m in a room with short reverberation time, present significant differences to those of the same signal recorded in 1 m at a large room with longer RT.

TABLE VI
PERFORMANCE OF THE METHOD FOR (a) ROOM A. THE ROWS REPRESENT THE ACTUAL CLASSES (a) AND THE COLUMNS THE PREDICTED CLASSES (p). HERE THE SYSTEM IS TESTED FOR VARIOUS DISTANCES THAT IT HAS NOT BEEN TRAINED FOR

A/P	0 m	0.5 m	1 m	2 m	3 m
0 m	99	1	0	0	0
0.5 m	2	97	0	0	1
1.5 m	0	0	25	39	36
2.5 m	0	0	10	39	51
3.5 m	2	0	6	22	70

2) Distance Mismatch: This experiment was conducted in order to examine how the method performs when it is tested within a single room, but for distances that were not included during the training. For this, the system was trained in Room A using the distances 0 m, 0.5 m, 1 m, 2 m, 3 m and then its performance was evaluated in the same room but for distances 0 m, 0.5 m, 1.5 m, 2.5 m, 3.5 m. The results of the test are given in Table VI. It can be seen that distances 0 m and 0.5 m were classified in the same way as in confusion matrix of Table III(a). This was expected, because the system has been trained with those two distances. The distance of 1.5 m was classified among the classes of 1 m (25%), 2 m (39%), and 3 m (36%). In the case of the distance of 2.5 m, 39% of the cases were classified at 2 m and 51% of the cases at 3 m. Finally, the distance of 3.5 m was classified for 70% of the cases as a distance of 3 m. This test indicates that for distances that the system has not been trained for, most likely a decision will be made assigning the nearest distance class employed during training.

3) Position Mismatch: This experiment evaluates the performance of the method, when there is a small mismatch ($\approx$ 10 cm offset) from the initial training classes. As shown in Fig. 9, the experiment took place in Room A and the system was trained at distances 0 m, 0.5 m, 1 m, 2 m, and 3 m. At the first test (Test A) the method was evaluated at distances 0 m, 0.4 m, 0.9 m, 1.9 m, and 2.9 m from the source. At the second test (Test B) the method was tested at 0 m, 0.6 m, 1.1 m, 2.1 m, and 3.1 m. Finally, during the third and fourth test the receiver was placed either at 10 cm to the right (Test C) or at 10 cm to the left (Test D) from the original training positions (see Fig. 9). For all the above test cases, the performance of the method was found to be reduced by no more than 2%, indicating the robustness of the method for small position mismatches.

V. COMPARISON TO EXISTING METHODS

Since there are no known earlier publications on single channel distance estimation directly from speech signals, it is not feasible to assess the relative performance of the proposed method. Nevertheless, in this section, the proposed method is compared to two other existing distance detection techniques [14], [28] that employ binaural signals. The first method [28] [Binaural fast Fourier transform (FFT)] is based on the logarithmic ratios of the Fourier transforms of the left and right signals, while the second work (Binaural MSC) uses the frequency dependent magnitude squared coherence (MSC) [14]. The results for the binaural methods were not obtained using our own measurements, but by summing up the reported results from [14], indicating that the existing state of the art binaural distance estimation methods perform much better than the proposed monoaural method.

TABLE VII
COMPARISON OF THE PROPOSED METHOD TO TWO OTHER BINAURAL METHODS

Method	Performance	Perf.(mismatch)
Proposed method (RT=0.39 s)	75.4%	73%
Binaural FFT (RT=0.3 s, 0.6 s)	97.9%	18.8%
Binaural MSC (RT=0.3 s, 0.6 s)	96.4%	68.8%

Table VII shows the performance of the proposed method compared to these comparison methods and it can be seen that these methods achieve more than 20% higher performance than the proposed method. However, for the case of a small position mismatch it can be seen that the Binaural FFT method is very sensitive and fails to classify, while the Binaural MSC method's performance decreases, but to a smaller extent. Interestingly, the proposed method presents better robustness to such small position mismatches. Furthermore, it should be noted that the performance of the two comparison methods was calculated by taking the mean of the results for two rooms having a RT of 0.3 s and 0.6 s. The proposed method was tested in a room with a RT of 0.39 s and it is expected that the performance of the proposed method would possibly further decrease for a room with RT of 0.6 s, as described in Section IV-D. Moreover, the two comparison methods employ not only distance estimation, but also angle detection and they are able to recognize distance using much shorter window lengths compared to the proposed method.

VI. LISTENING TEST

In order to compare the results of the proposed method to the performance of human listeners for automatically detecting distance, a listening test was conducted using the same data and settings as for the cases described in Section IV. In each run, the test subjects were asked to detect the distance of speech signals recorded at different distances from the source. The subject's task was to assign one value (0, 0.5, 1, 2, 3) to each speech signal and their choices should correspond to the apparent perceived distance of the sound source in meters (0 m, 0.5 m, 1 m, 2 m, 3 m). The same database, described in Section IV-A was used and the listening test consisted of four different sessions, one for each room of Table I. Before each session the test subjects had the opportunity to listen to speech signals recorded at several distances from the source. They could choose a distance value and then listen to the speech signal recorded at the corresponding selected position in the specific room and in this way, they were able to obtain an impression of the acoustical properties of the room. In this sense, this training session can be considered to be comparable to the training stage of the classifier. The recordings were normalized to a maximum value of 0-dB Full Scale (FS) and they were presented with headphones monaurally to the test subjects. Ten normal hearing subjects participated in the experiment. The mean performance of human listeners is presented in Fig. 10, using errorbars to indicate the standard error from the mean. It can be seen that there is large variability of the responses between individual listeners. This variability is in agreement with [50] and as is suggested [51] this is primarily due to perceptual blur in the auditory domain.

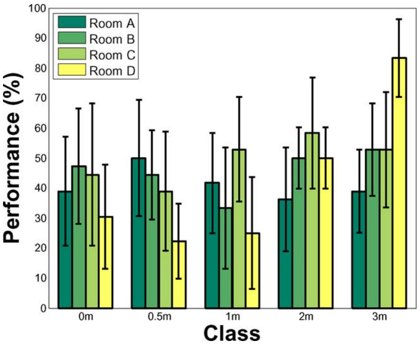


Fig. 10. Results obtained from the listening test for (a) Room A, (b) Room B, (c) Room C, (d) Room D.

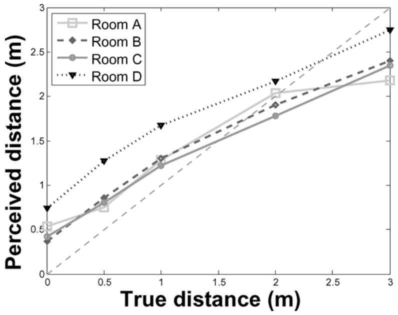


Fig. 11. Perceived distance as a function of true distance for human listeners. Close source distances are overestimated and longer distances are underestimated.

In Fig. 11, perceived distance is plotted as a function of the true distance. It can be seen that, on average, monoaural human performance is not as good as the performance of the proposed method. For small distances, the perceived distance is almost proportional to source distance and increases slowly, when source distance is more than 2 m. This effect has also been found in earlier studies of distance perception [18], [19], [25], [52]. Additionally, it can be noted that close source distances are overestimated and longer distances are underestimated. This effect might be related to the auditory horizon that represents the maximum perceived distance [53].

Moreover, from Fig. 11 it can be seen that the perceived distance depends also on the reverberation time of the room, which is also in agreement with [52].

VII. DISCUSSION AND CONCLUSION

The proposed system employs a novel methodology for detecting the distance of any speaker using a single microphone receiver. Several features based on the spectral and temporal characteristics of speech have been examined and proved to be dependent on the distance between source and receiver inside typical reverberant rooms. These features were derived in such a way that they are independent of the signal gain and level, thus are not affected by the individual speaker output level and microphone setup.

The robustness of the method was found to depend on the reverberation time of the room and the longer the reverberation time, the lower the performance. The method was tested for both speaker-dependent and speaker-independent conditions and it was found that for rooms with low reverberation there was a small increase in performance when the system was trained and tested with the same speakers. In the case of the rooms with longer reverberation time, it was found that the specific speaker is not critical as the room acoustics have much stronger effect on the received signals and on the accuracy of the method.

It was also observed that the performance of the method was significantly lower for larger distances (2–3 m). That is probably related to the relative dominance of the reverberant sound field for these distances, which presumably does not depend a lot on distance, resulting in small changes in feature values for these distances.

The choice of the block duration used for the feature extraction was also examined and it was shown that for all the rooms apart from the room with the shortest reverberation time, performance increases significantly for longer blocks. However, increasing block size introduces latency and restricts the range of potential applications.

This method is robust to small position mismatches, but it is sensitive to the RT of the room and it is necessary to train the system in a room with similar acoustical properties to the room where the system is used. Moreover, as the system is trained with specific distances, when it is tested for distances that it has not been trained for, a likely decision is made to the nearest distance employed during training.

Overall, the proposed method provides a good distance detector, especially for smaller distances and the method is specific for speech signals, but is not speaker specific. Its overall performance may be lower when compared to binaural methods, but nevertheless it appears to be robust to small distance mismatches.

In contrast, tests conducted with human listeners under identical conditions indicate that there is a large variability of the performance between individual listeners and that the mean human performance is lower than that of the developed classifier.

The method presented allows for the estimation of distance based on single microphone signals and gives best performance for close distances. This makes this method useful for an ambient telephony system or a distributed sensor system, where each terminal can make an independent estimate of the distance to the speaker. The closest terminal can then be selected to capture and reproduce the signals with relatively little network communication load.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and constructive suggestions to improve this paper.

References

[1] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, “Signal processing in high-end hearing aids: State of the art, challenges, and future trends,” *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2915–2929, 2005.

[2] M. Omologo, P. Svaizer, and M. Matassoni, “Environmental conditions and acoustic transduction in hands-free speech recognition,” *Speech Commun.*, vol. 25, no. 1-3, pp. 75–95, Aug. 1998.

[3] *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

[4] *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. New York: Wiley-IEEE, 2006.

[5] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, “Augmented reality audio for mobile and wearable appliances,” *J. Audio Eng. Soc.*, vol. 52, pp. 618–639, Jun. 2004.

[6] S. Oh, V. Viswanathan, and P. Papamichalis, “Hands-free voice communication in an automobile with a microphone array,” in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Los Alamitos, CA, 1992, vol. 1, pp. 281–284.

[7] S. Gustafsson, R. Martin, and P. Vary, “Combined acoustic echo control and noise reduction for hands-free telephony,” *Signal Process., Special Iss. Acoust. Echo Noise Control*, vol. 64, no. 1, pp. 21–32, 1998.

[8] A. Härmä, “Ambient human-to-human communication,” in *Handbook of Ambient Intelligence and Smart Environments*. New York: Springer, 2009, pp. 795–823.

[9] A. Härmä, “Ambient telephony: Scenarios and research challenges,” in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[10] A. Härmä, S. van de Par, and W. de Bruijn, “Spatial audio rendering using sparse and distributed arrays,” in *Proc. 122nd AES Conv.*, Vienna, Austria, May 2007.

[11] A. Härmä and K. Pham, “Conversation detection in ambient telephony,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4641–4641.

[12] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[13] *Microphone Array Signal Process*, J. Benesty, J. Chen, and Y. Huang, Eds. Berlin, Heidelberg, Germany: Springer, 2008.

[14] S. Vesa, “Binaural sound source distance learning in rooms,” *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

[15] Y. C. Lu and M. Cooke, “Binaural distance perception based on direct-to-reverberant energy ratio,” in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2008.

[16] P. Zahorik, S. D. Brungart, and W. A. Bronkhorst, “Auditory distance perception in humans: A summary of past and present research,” *Acta Acustica*, vol. 91, pp. 409–420, May/Jun. 2005.

[17] P. Zahorik, “Direct-to-reverberant energy ratio sensitivity,” *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 2110–2117, 2002.

[18] D. H. Mershon and J. N. Bowers, “Absolute and relative cues for the auditory perception of egocentric distance,” *Percept.*, vol. 8, pp. 311–322, 1979.

[19] D. H. Mershon and E. King, “Intensity and reverberation as factors in auditory perception of egocentric distance,” *Percept. Psychophys.*, vol. 18, no. 6, pp. 409–415, 1975.

[20] P. Zahorik, “Assessing auditory distance perception using virtual acoustics,” *J. Audio Eng. Soc.*, vol. 111, pp. 1832–1846, 2002.

[21] N. Sakamoto, T. Gotoh, and Y. Kimura, “On “out-of-head localization” in headphone listening,” *J. Audio Eng. Soc.*, vol. 24, pp. 710–716, 1976.

[22] D. R. Begault, “Perceptual effects of synthetic reverberation on three-dimensional audio systems,” *J. Audio Eng. Soc.*, vol. 40, pp. 895–904, 1992.

[23] R. A. Butler, E. T. Levy, and W. D. Neff, “Apparent distance of sounds recorded in echoic and anechoic chambers,” *J. Experim. Psychol.*, vol. 6, pp. 745–750, 1980.

[24] J. W. Philbeck and D. H. Mershon, “Knowledge about typical source output influences perceived auditory distance,” *J. Audio Eng. Soc.*, vol. 111, pp. 1980–1983, 2000.

[25] S. H. Nielsen, “Auditory distance perception in different rooms,” *J. Audio Eng. Soc.*, vol. 41, pp. 755–770, 1993.

[26] A. W. Bronkhorst, “Localization of real and virtual sound sources,” *J. Audio Eng. Soc.*, vol. 98, pp. 2452–2553, 1995.

[27] S. Vesa, “Sound source distance learning based on binaural signals,” in *Proc. Workshop Applicat. Signal Process., Audio, Acoust. (WASPAA’07)*, 2007, pp. 271–274.

[28] P. Smaragdis and P. Boufounos, “Position and trajectory learning for microphone arrays,” *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 15, no. 1, pp. 358–368, Jan. 2007.

[29] Y.-C. Lu, M. Cooke, and H. Christensen, “Active binaural distance estimation for dynamic sources,” in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[30] N. Lesser and D. Ellis, “Clap detection and discrimination for rhythm therapy,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 3, pp. 37–40.

[31] T. Takiguchi, Y. Sumida, and Y. Ariki, "Estimation of room acoustic transfer function using speech model," in *Proc. IEEE/SP 14th Workshop Statist. Signal Process.*, Los Alamitos, CA, 2007, pp. 336–340.

[32] A. Saxena and A. Ng, "Learning sound location from a single microphone," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Kobe, Japan, May 2009.

[33] A. Vähätalo and I. Johansson, "Voice activity detection for GSM adaptive multi-rate codec," in *Proc. IEEE Workshop Speech Coding Process.*, 1999, pp. 55–57.

[34] M. Shashanka, B. Shinn-Cunningham, and M. Cooke, "Effects of reverberant energy on statistics of speech," in *Proc. Workshop Speech Separation Comprehension Complex Acoust. Environ.*, Montreal, QC, Canada, Nov. 2004.

[35] E. Georganti, J. Mourjopoulos, and F. Jacobsen, "Analysis of room transfer function and reverberant signal statistics," in *Proc. Acoust.'08*, Paris, France, 2008.

[36] E. Georganti, T. Zarouchas, and J. Mourjopoulos, "Reverberation analysis via response and signal statistics," in *Proc. 128th AES Conv.*, London, U.K., 2010.

[37] B. Gillespie, D. A. F. Florencio, and H. S. Malvar, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 3701–3704.

[38] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[39] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[40] D. Fee, C. Cowan, S. Bilbao, and I. Ozcelik, *Predictive Deconvolution and Kurtosis Maximization For Speech Dereverberation*. Florence, Italy: , 2006.

[41] M. Wu and D. Wang, "Two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[42] K. Furuya, S. Sakauchi, and A. Kataoka, "Speech dereverberation by combining mint-based blind deconvolution and modified spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, vol. 1, pp. 813–816.

[43] A. Tsilfidis and J. Mourjopoulos, "Signal-dependent constraints for perceptually motivated suppression of late reverberation," *Signal Process.*, vol. 90, pp. 959–965, Mar. 2010.

[44] T. Zarouchas and J. Mourjopoulos, "Modeling perceptual effects of reverberation on stereophonic sound reproduction in rooms," *J. Acoust. Soc. Amer.*, vol. 126, pp. 229–242, Jul. 2009.

[45] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[46] J. J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoust. Soc. Amer.*, vol. 61, no. S1, pp. S34–S34, 1977.

[47] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York: Springer, 2006.

[48] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[49] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[50] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Amer.*, vol. 112, pp. 2110–2117, 2002.

[51] J. M. Loomis, J. W. Philbeck, and P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 28, pp. 1202–1212, 2002.

[52] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, 1999.

[53] G. von Békésy, "The moon illusion and similar auditory phenomena," *Amer. J. Psychol.*, vol. 111, pp. 1832–1846, 2002.

Eleftheria Georganti received the diploma degree from the Department of Electrical Engineering and Computer Engineering, University of Patras, Rio. Greece, in 2007. She is currently pursuing the Ph.D. degree at the University of Patras, with a thesis on "Modeling, analysis, and processing of room transfer functions under reverberant condition" at the Audio and Acoustic Technology group of the Wire Communications Laboratory.

She carried out nine months of her research at the Technical University of Denmark (DTU) under the framework of Marie Curie Host Fellowships for Early Stage Research Training (EST) and another nine months at the DSP Group of Philips Research, Eindhoven, The Netherlands, working on ambient telephony technologies. Her research interests include acoustical room responses modeling, psychoacoustics, and statistical signal processing.

Tobias May received the Dipl.-Ing. (FH) degree in hearing technology and audiology from the Oldenburg University of Applied Science, Oldenburg, Germany, in 2005 and the M.Sc. degree in hearing technology and audiology from the University of Oldenburg, Oldenburg, Germany, in 2007. He is currently pursuing the Ph.D. degree at the University of Oldenburg.

Since 2007, he has been with the Eindhoven University of Technology, Eindhoven, The Netherlands. Since 2010, he has been affiliated with the University of Oldenburg. His research interests include computational auditory scene analysis, binaural signal processing, and automatic speaker recognition.

Steven van de Par studied physics at the Eindhoven University of Technology, Eindhoven, The Netherlands, and received the Ph.D. degree from the Eindhoven University of Technology in 1998 on a topic related to binaural hearing.

As a Postdoctoral Researcher at the Eindhoven University of Technology, he studied auditory-visual interaction and was a Guest Researcher at the University of Connecticut Health Center. In early 2000, he joined Philips Research, Eindhoven, to do applied research in digital signal processing and acoustics. His main fields of expertise are auditory and multisensory perception, low-bit-rate audio coding, and music information retrieval. He has published various papers on binaural auditory perception, auditory–visual synchrony perception, audio coding, and music information retrieval (MIR)-related topics. Since April 2010, he has held a professor position in acoustics at the University of Oldenburg, Oldenburg, Germany.

Aki Härmä received the Ph.D. degree from the Helsinki University of Technology, Espoo, Finland, in 2001 on frequency-warped signal processing algorithms.

In 2000–2001, he was a Consultant at Lucent Bell Laboratories and Agere Systems, Murray Hill, NJ. In 2001, he returned to the Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, and since 2004 he has been with the Digital Signal Processing Group of Philips Research Laboratories, Eindhoven, The Netherlands. His main research interests are in the areas of acoustics, speech, and audio signal processing, pattern recognition, source separation, communication, perception, and user interaction.

John Mourjopoulos (M'90) received the B.Sc. degree in engineering from Coventry University, Coventry, U.K., in 1978, the M.Sc. and Ph.D. degrees from the Institute of Sound and Vibration Research (ISVR), Southampton University, Southampton, U.K., in 1980 and 1985, respectively.

Since 1986, he has been with the Electrical and Computing Engineering Department, University of Patras, Rio, Greece, where he is now Professor of Electroacoustics and Digital Audio Technology and head of the Audio and Acoustic Technology Group of the Wire Communications Laboratory. In 2000, he was a Visiting Professor at the Institute for Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. He has authored and presented more that 100 papers in international journals and conferences. He has worked in national and European projects, has organized seminars and short courses, has served in the organizing committees and as session chairman in many conferences, and has contributed to the development of digital audio devices. His research covers many aspects of digital processing of audio and acoustic signals, especially focusing on room acoustics equalization. He has worked on perceptually motivated models for such applications, as well as for speech and audio signal enhancement. His recent research also covers aspects of the all-digital audio chain, the direct acoustic transduction of digital audio streams, and WLAN audio and amplification.

Prof. Mourjopoulos was awarded the Fellowship of the Audio Engineering Society (AES) in 2006. He is a member of the AES (currently serving as section vice-chairman) and of the Hellenic Institute of Acoustics being currently its vice-president.