Microphone Arrays and Speaker Identification

Qiguang Lin, *Member, IEEE,* Ea-Ee Jan, *Student Member, IEEE,* and James Flanagan, *Life Fellow, IEEE*

***Abstract*— Hands-free operation of speech processing equipment is sometimes desired so that the user is unencumbered by hand-held or body-worn microphones. This paper explores the use of array microphones to capture speech under adverse acoustic conditions, and provide input to a system for automatic speaker identification. The system is evaluated using reverberated speech signals, generated by a computer model of room acoustics and transduced by different simulated microphone-arrays. For comparison, the system is also evaluated using close-talking microphone input. It is found that 2-D matched-filter microphone arrays are capable of producing high speaker identification scores in a hostile acoustic environment, such as multipath distortion and competing noise sources. The paper also explores the influence of vector quantization techniques, codebook size, and order of cepstrum coefficients on the performance of the speaker identification system.**

I. Introduction

PERFORMANCE of speech and speaker recognizers is typically degraded by deleterious properties of the acoustic environment, such as multipath distortion (reverberation) and ambient noise. The degradation becomes more prominent as the microphone is positioned more distant from the speaker, for instance, in a teleconferencing application. Previous work has demonstrated that beamforming/matched-filter microphone arrays can provide higher signal-to-noise ratios (SNR) than can conventional microphones used at distances (see e.g., [1]). Consequently, there is increasing interest in microphone arrays for hands-free operation of speech processing systems.

In this paper, we study speech enhancement by microphone arrays as applied to speaker identification. Speaker identification is chosen for experiment because its classifier is generally simpler than that used in speech recognition. That is, experimental results are less likely to be influenced by details of the back end, such as speaker dependency and language models. These factors are known to be significant in speech recognition. Furthermore, since our speaker identification system operates in a text-independent mode, it relies exclusively on the acoustic information carried in individual frames of the time waveform. Therefore, we can more reliably evaluate the capability of microphone arrays for speech enhancement. An additional incentive is that we want to develop a speaker identification system for multimedia environments. The capability for identification and separation of multiple talkers is important in conferencing applications [1]. Our work therefore focuses on speaker identification from speech signals degraded by room reverberation and/or competing noise sources, using microphone-arrays as sound capture equipment.

This paper is organized as follows. Section II describes a vector-quantization (VQ) based, text-independent speaker identification system. This section addresses the effects of various VQ techniques, codebook sizes, and order of cepstrum coefficients on the performance of speaker identification. Section III summarizes a computer model for simulation of room acoustics and principles of matched-filter arrays. Section IV describes generation of degraded speech data for use in the system evaluation. Section V discusses evaluation results for both matched and unmatched testing and training conditions. Finally, Section VI gives general conclusions.

Manuscript received December 21, 1993; revised April 5, 1994. This work was initiated during the 1993 DOD Speech Workshop at the CAIP Center, Rutgers University, supported by National Security Agency Contract MDA904-93-4073. Post-workshop research has been supported by ARPA Contract DABT63-93-C-0037 and NSF Grant MIP-9121541.

The authors are with the Center for Computer Aids for Industrial Productivity, Rutgers University, Piscataway, NJ 08855-1390 USA.

IEEE Log Number 9403965.

II. Speaker Identification

Speaker identification, verification, and classification are three aspects of speaker recognition. The task of speaker identification is to automatically determine whether an utterance has been spoken by one of a population of speakers and to identify the talker. If the talker is known to be represented in the population, the identification is from a "closed set." If not, it is from an "open set." The identification task is different from that of speech recognition. Speech recognition is an automatic extraction of what has been said. Often, personal identity information is not of concern or is "neutralized," e.g., in a speaker-independent speech recognition system. Despite the inherent differences between speaker identification and speech recognition, both applications are based on similar front-end acoustic analysis techniques and speech features. For instance, linear prediction coding (LPC) derived cepstrum coefficients have been widely used both for speaker identification and speech recognition. The power of cepstrum representation of the short-time speech spectrum lies in the fact that the L2-norm can be utilized as the distance measure and that cepstral basis functions are orthogonal [2].

Speaker recognition is typically divided into two subclasses, according to text dependency: (1) text-dependent and (2) text-independent. In the text-dependent mode, speakers provide utterances of the same text for both training and testing. A classical approach to text-dependent speaker recognition is template matching or pattern recognition, where dynamic time-warping methods are usually applied to temporally align the training and testing utterances [3]–[7]. Further, because the same sequence of phonetic events are spoken, dynamic characteristics of spectra (such as delta cepstrum) are often included in the feature representation to improve recognition. Text-independent speaker recognition does not impose the con-

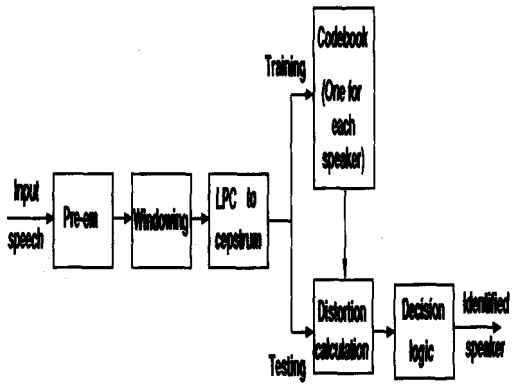


Fig. 1. Block diagram of the vector-quantization-based speaker identification system.

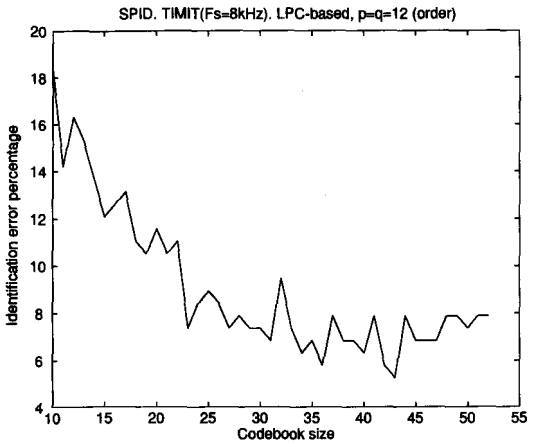


Fig. 2. Speaker identification error percentage as a function of LVQ codebook size. $p = 12$ is the order of LPC and $q = 12$ is the order of cepstrum coefficients.

straint of the same text in training and recognition trials. Consequently, text-independent speaker recognition techniques are primarily based on measurements without reference to a timing index [8], [9], and hence, dynamic attributes of spectra cannot be fully exploited.

In this study, a closed-set, text-independent speaker identification system is developed. The system uses LPC-derived cepstrum coefficients as the measured features and follows speaker-based VQ approaches [8]. Fig. 1 shows a block diagram of the identification system. During training sessions, the input speech signal is preemphasized, windowed, and LPC analyzed, resulting in a sequence of vectors of LPC-derived cepstrum coefficients. The resultant cepstrum vectors are vector-quantized, with one codebook for each individual speaker. During testing trials, corresponding cepstrum coefficients of the test sentence are computed. The following distance, with respect to the ith codebook, is then determined

$$D^{(i)} = \frac{1}{L}\sum_{l=1}^{1} \min_{1\le m\le M} \left\{ \sum_{n=1}^{q} [c(l,n) - C_i(m,n)]^2 \right\} \quad (1)$$

where L is the number of frames in a test sentence, q is the order of cepstrum coefficients, M is the number of entries of the codebook, $c(l,n)$ is the cepstrum-coefficient vector of the testing sentence at the frame l, and $C_i(m,n)$ is the cepstrum vector of the mth centroid of the ith codebook. Note that the $zero$th-order cepstrum coefficient, which is a meausre of gain, is not used in (1). The final speaker identification decision is given by (see also [8])

$$i^* = \text{ argmin } D^{(i)}, \text{ for } 1 \le i \le I \quad (2)$$

where I is the total number of speakers in the closed set.

Proper design of the speaker-based codebook is important to the performance of speaker identification systems [8], [12]. How to choose different VQ algorithms and how to determine the codebook size M (i.e., the number of codebook entries)? In the remaining part of this section, the New England subset of the TIMIT database is used to quantify the effects of VQ algorithms, codebook size, order of cepstrum coefficients, threshold for silence removal, and frame size on system performance.

The TIMIT speech data are recorded from a close-talking microphone, bandlimited to 8 kHz and sampled at 16 kHz. The New England subset comprises 24 males and 14 females. The TIMIT data are divided into a training set and a testing set. The training set for each speaker comprises ten sentences and these sentences are used for the present evaluation. The first five sentences are concatenated and used as training data for the speaker identification. The purpose of concatenation is to have more speech material to train the codebook. The remaining five sentences are used to test the speaker identification. Thus, the total number of trials is 190 (5×38). For our purposes, the TIMIT data are low-pass filtered to 4 kHz and down sampled at 8 kHz, to be representative of telephone bandwidth.

Fig. 2 shows identification error pecentage as a function of codebook size M. The learning vector quantization (LVQ) [10] is used as the vector quantizer. The order of LPC analysis and the order of cepstrum coefficients are both set to 12. A Hamming window of 15 ms is used and the window is stepped in time every 7.5 ms. The choice of the window length is based on a series of preliminary trials.

Fig. 2 shows that a codebook size of $M = 43$ yields the smallest error rate (5.3%), though the curve shows certain fluctuations. A general trend of Fig. 2 is that the error rate decreases when the codebook size M increases up to 43. For $44 \le M \le 52$ the error rate is in the range of 7%–8%.

Another commonly used vector quantizer is the binary-splitting algorithm of [11], henceforth referred to LBG. The LBG method is easy to implement, but it has a constraint that the codebook size must be an integer power of 2. In Fig. 3, identification error rates obtained using LVQ and LBG are plotted for codebook sizes of $M = 16, 32, 64$, and 128. The results indicate that LBG performs slightly better than LVQ for $M = 32$ and 64. However, LBG could not reach the

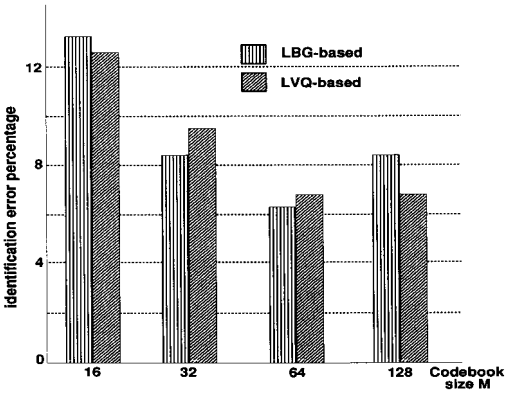


Fig. 3. Speaker identification error percentage for LVQ and LBG vector quantizers.

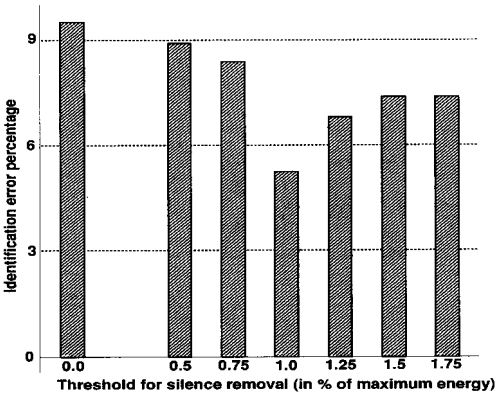


Fig. 4. Speaker identification error percentage as a function of threshold for silence removal.

TABLE I
SPEAKER IDENTIFICATION ERROR PERCENTAGE AS A FUNCTION OF CODEBOOK SIZE M, LPC ORDER p, CEPSTRUM ORDER q, FRAME LENGTH N, AND FRAME PERIOD $K \cdot N$ AND K ARE IN NUMBER OF SAMPLES AND THE VECTOR QUANTIZER USED IN LVQ. THE CODEBOOK SIZES OF $M = 38$ AND $M = 46$ ARE CHOSEN TO COMPARE WITH AVAILABLE DATA [12].

Codebook size M	*N=120, K=60*			*N=240, K=60*	*N=240, K=120*
	p=q=12	*p=12, q=18*	*p=q=18*	*p=q=12*	
38	6.8	8.9	4.7	7.4	10.5
43	5.3	5.8	7.4	10.0	8.4
46	6.8	6.8	6.8	8.9	10.5

minimum error percentage obtained by LVQ for a codebook size of 43.

Use of higher-order cepstrum coefficients does not necessarily improve the identification score for a given codebook size. Table I gives identification results for different cepstrum orders, q. In this experiment, higher-order cepstrum coefficients are obtained either by additionally computing higher-order LPC coefficients or by setting the higher LPC coefficients to zero. As exemplified in Table I, the relationship between cepstrum order and identification error is not straightforward. Table I also illustrates the effect of frame length on identification scores. It should be mentioned again that the present speaker identification system, which is text-independent, does not incorporate delta cepstrum coefficients.

Optimal choice of codebook size, frame length, LPC order, and cepstrum order is therefore a multidimensional problem. It is difficult to find the global optimum, because the solution may depend on the speech data and because some of the differences may not be statistically significant. In addition, the LVQ algorithm may converge to a local minimum, instead of the global minimum. We therefore adopt the condition used for Fig. 2 as the reference in subsequent tests, with the codebook size being fixed at $M = 43$. The LPC order and the cepstrum order are both set to 12.

Another important factor which influences the performance of speaker identification is the threshold for silence removal. The purpose of silence removal is to eliminate segments in which only background noise is presented. Soong *et al.* [8] found that use of only voiced frames consistently led to a poorer result than use of all speech frames. These authors ascribed the degradation to inconsistency between training and testing, because unvoiced frames were not deliberately removed during training. In the present study, a silence-removal routine is included both in training and in testing. The effect of the threshold setting can be seen in Fig. 4.

The results of Fig. 4 can be interpreted as follows. In VQ approaches, where no explicit phoneme classification is made, beginning/ending silence-segments tend to degrade speaker identification performance. On the other hand, a too high threshold will not only remove silence segments, but also transitional segments between phonemes. Transitional segments may convey speaker-specific information, and hence, can reduce identification errors if they are included in calculating distortions. Moreover, as the threshold increases, more frames will be disregarded and the effective amount of testing data as well as training data decreases. A short test sentence tends to give a poorer result than a long sentence does, especially for text-independent speaker identification.

III. COMPUTER MODEL OF ROOM ACOUSTICS AND MATCHED-FILTER MICROPHONE ARRAYS

A. Image Model

A typical onmidirectional microphone is sensitive to pressure fluctuations in an acoustic wave. For a point sound source of sinusoidal frequency ω, the temporal variation of sound pressure at r-distance from the source is

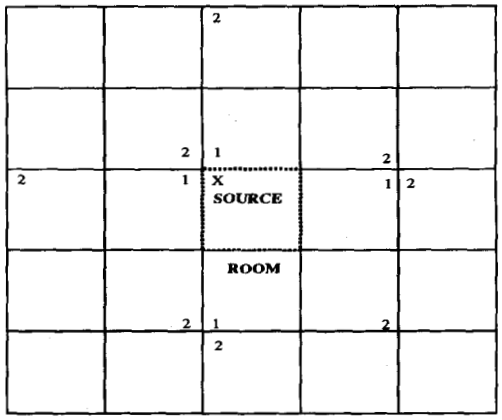
$$p(r,t) = \frac{A}{r} e^{j\omega(t-r/c)} \tag{3}$$

where A is the source strength and c is the speed of sound.

The walls of conventional rooms are large and the roughness dimensions of the wall surfaces are small compared to acoustic wavelengths of interest. Therefore, the walls constitute effective reflectors (mirrors) and acoustic wave propagation, from source to receiver, can be determined by ray tracing

Order of Images	Number of Images	Cumulative Total Number of Images
0	1	1
1	6	7
2	18	25
3	38	63
4	66	129
5	102	231

(a)



(b)

Fig. 5. (a) Number of acoustic images and cumulative total number of images associated with order for 3-D simulation of a room. (b) The diagram illustrates the location of images of order 1 and 2 corresponding to a sound source located a X for a 2-D enclosure. The dashed lines represent the walls of the room.

techniques. Using principles described by Allen and Berkley [13], Jan and Flanagan [14] have implemented a computer model of room acoustics, based on the image technique of computing source-receiver impulse responses. Image sources are determined in accordance with Snell's Law. The strength of the images depends upon the absorption of the walls that each ray path encounters. For simplicity, the absorption is assumed to be frequency independent in the model. The sound pressure contributed by each source is given by (3) and includes spherical spreading in the enclosure. The model does not account for detailed scattering and diffusion effects.

Fig. 5 shows an image diagram in 2-D for a hard-walled room, as well as the number of images through fifth order for a 3-D rectangular enclosure. One notices that the total number of images for a given order of reflection is not a geometric series, but rather resembles an arithmetic series. Refer to [14] for details.

B. Matched-Filter Array

Matched-filter techniques can be applied to microphone arrays to improve noise rejection and achieve spatial selectivity in 3-D. A matched-filter is the time inverse of the impulse response of the system to be matched. In the array, a matched-filter is dedicated to each microphone. The array output is the summation of outputs from each matched-filter. Because the time-inverse impulse response, $h(-t)$, is typically noncausal,

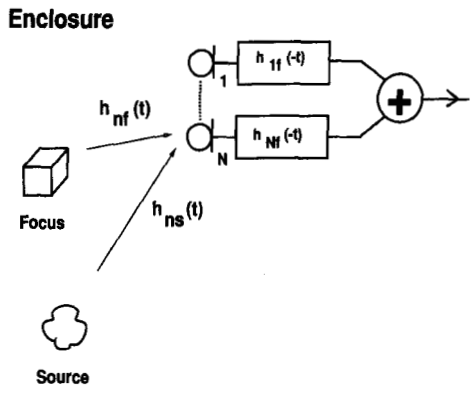


Fig. 6. Configuration of a matched-filter microphone array with N microphones. An off-focus condition is shown for the sound source.

a truncation and fixed delay are required to realize a causal filter which approximates the desired response.

The impulse response from the desired focal point to each receiver in the array is required to implement the matched-filter array system. This response can be calculated from the room geometry or measured in actual rooms. For a source located at the focal point emitting a signal $s(t)$, the temporal output of the matched-filter array is

$$O_f(t) = \sum_{n=1}^{N} s(t) * h_{nf}(t) * h_{nf}(-t)$$
$$= s(t) * \sum_{n=1}^{N} h_{nf}(t) * h_{nf}(-t) \qquad (4)$$

where $h_{nf}(t)$ is the impulse response from the focal point to the nth sensor, N is the total number of sensors, and $*$ denotes convolution. The term denoted by the summation in (4) is recognized as the autocorrelation of the impulse response from focus to sensor. When the source is off the focal position, as shown in Fig. 6, the temporal output of the array is

$$O_o(t) = s(t) * \sum_{n=1}^{N} h_{ns}(t) * h_{nf}(-t) \qquad (5)$$

where $h_{ns}(t)$ is the impulse response from the source to the nth sensor. In (5), the term denoted by summation is recognized as the cross-correlation of the impulse responses from focus to sensor and from source to sensor. One sees that the size of the focal volume for retrieval of low distortion signals is conditioned by the spatial correlation of the impulse responses $h_{nf}(t)$ and $h_{ns}(t)$.

Fig. 7 illustrates the spatial correlation of the system impulse response, $\phi(h_{nf}, h_{ns}) = \sum_{n=1}^{N} h_{ns}(t) * h_{nf}(-t)$. Two 31×31 orthogonal arrays are used and placed in the center of adjacent walls in a $20 \times 16 \times 3$ m room. The arrays are focused at (14, 9.5, 1.7) m which is identical to the location of the signal. That is, the source is on focus. Fig. 8 shows the spatial correlation function $\phi(h_{nf}, h_{ns})$ for an off-focus source. Spatial volume

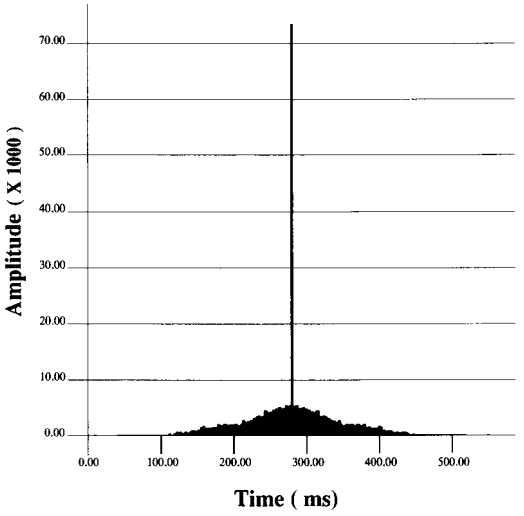


Fig. 7. Impulse response for two orthogonal 31 × 31 matched-filter arrays at a focused condition in a room of dimensions of 20 × 16 × 3 m. The wall absorption is 0.1 and images up to fifth order are included. The system is focused at (14, 9.5, 1.7) m, coinciding with the source position.

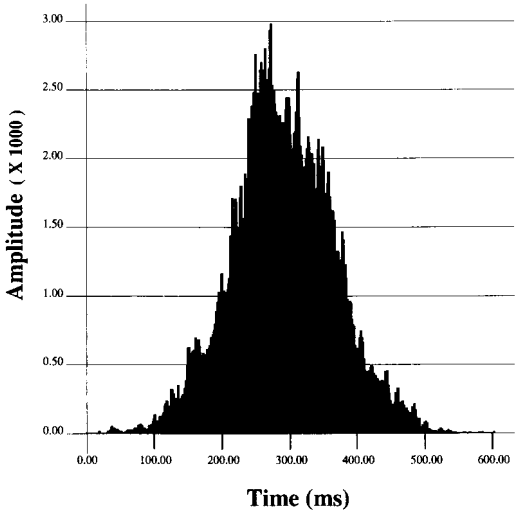


Fig. 8. Impulse response for two orthogonal 31 × 31 matched-filter arrays when the microphone system is focused at (14, 9.5, 1.7) m and the source is at (5.0, 3.0, 1.0) m. Other conditions are the same as in Fig. 7.

selectivity of the matched-filter arrays is implied by Figs. 7 and 8.

IV. Generation of Degraded Speech

The room simulation is used to generate speech degraded by a multipath environment for evaluation of our speaker identification system. Dimensions of the simulated enclosure are 20 × 16 × 5 m. The acoustic absorption coefficient for all walls is set to 0.1 producing a highly reverberant enclosure. Images up to fifth order are included in generating the reverberant speech. The reverberation time of the enclosure is approximately 1.6 s. Inputs to the room simulation are close-talking speech described in Section II (the TIMIT database). The signal source is placed at the point (14, 9.5, 1.7) m. A competing noise source of variable intensity, to produce SNR's can be turned on or off. The noise is generated by a Gaussian random number generator and is located at (3.0, 5.0, 1.0) m.

Four sound pickup systems are used to capture the degraded speech:

1) *Single microphone system.* A single microphone receiver is located at (10, 0.5, 1.7) m. The overall system impulse response is simply the impulse response from the source to the receiver. The microphone is an omnidirectional pressure sensor.
2) *One beamforming array.* A single-beam line array is placed 0.5 meter off the wall with its center located at (10, 0.5, 1.7) m. The array consists of 51 microphones with uniform separation of 4 cm. This spacing provides selectivity without spatial aliasing for frequencies up to 4000 Hz. The impulse response from the source to each receiver is calculated from the room geometry. The direct path-arrivals of the impulse responses are used to produce a single-beam delay-and-sum beamformer [1], [15].
3) *Two beamforming array.* Two orthogonal line arrays are placed 0.5 meter off the walls with their centers at (10, 0.5, 1.7) m and (0.5, 8.0, 1.7) m, respectively.
4) *Matched-filter arrays.* Two 31 × 31 matched-filter arrays are placed on orthogonal walls. Centers of the arrays are at (10, 0, 1, 7) m and (0, 8, 1.7) m, respectively. Separation of microphone elements is again 4 cm. The overall system impulse response is calculated from (4) with appropriate temporal alignment [1].

Indicative of the difference between close-talking speech and reverberant speech received by different microphone systems, signal waveforms are given in Fig. 9. The top panel is close-talking speech (from the TIMIT database), the middle panel is reverberant speech received by a single microphone, and the bottom is reverberant speech captured by the two 2-D matched-filter microphone arrays.

V. Identification Results

Speaker identification systems can be evaluated under two conditions: matched or unmatched conditions between training and testing. A matched condition corresponds to training and testing under identical system conditions, i.e., frequency response, microphone, SNR, reverberation, etc. An unmatched condition corresponds to training and testing with system differences. Generally, a matched condition gives higher identification scores than an unmatched condition.

A. Results for Matched Conditions of Training and Testing

Fig. 10 gives results of speaker identification for close-talking speech and distant-talking reverberant speech captured by four different microphone systems: A single omnidirec-

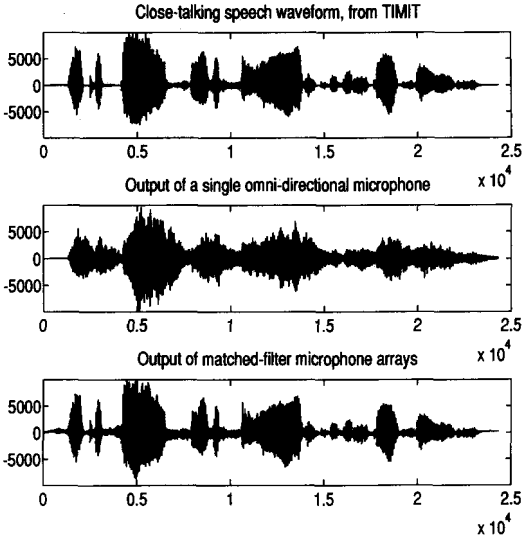


Fig. 9. Speech waveforms of close-talking (top), reverberant speech received with a single microphone (middle), and reverberant speech received with the matched-filter array microphones (bottom).

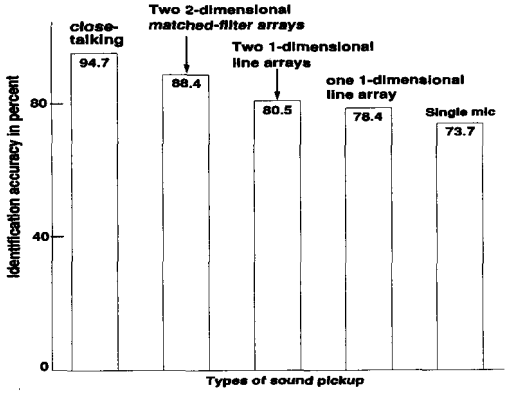


Fig. 10. Speaker identification under a matched training and testing condition. No competing noise source is presented and no silence removal is made. When "silence frames" of the reverberant speech are removed with a threshold of 2.5%, the matched-filter array system gives an identification score as high as 91.1%. Spatial locations of transducer and source are given in the text.

tional microphone, a 1-D line array, two 1-D line arrays, and two orthogonally-placed 2-D matched-filter microphone arrays.

As can be expected, close-talking speech gives the highest identification score 94.7%, followed by the reverberant speech captured by matched-filter arrays, 88.4%. The use of two 1-D line arrays increases the score by 2%, compared with that of one line array 78.4%. The single omnidirectional microphone gives the worst result of 73.7%. It is important to note that the threshold used for silence removal in Fig. 10 is all set to 1% of the maximum frame-energy. This value is essentially too low to eliminate any frames of the reverberant speech signals in

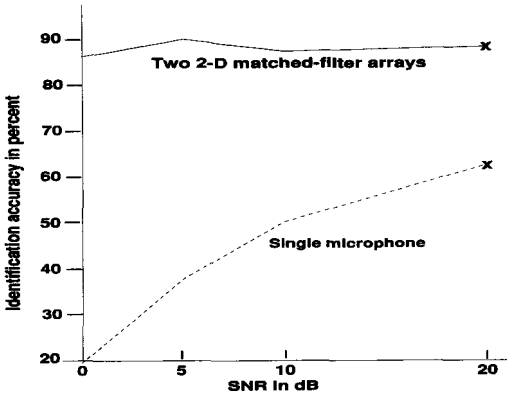


Fig. 11. Speaker identification under a matched training and testing condition. No silence removal is made. A competing noise is presented with varying SNR. The interfering noise SNR is calculated over the whole utterance and is defined as the ratio of the power of original speech signal to the power of the original noise signal. That is, the decrease in SNR due to room reverberation is not included. Reverberation distortion dominates when the intensity of the interfering noise is relatively low, marked with x in the diagram.

Fig. 10. If the threshold is increased to 2.5%, it is found that the matched-filter arrays gives a correct identification score as high as 91.1%. It is thus clear that the matched-filter microphone arrays can be used as high quality, hands-free sound pick-up. In the following discussion, we will accordingly focus on the difference in performance of speaker identification between the matched-filter arrays and the single microphone system.

The power of matched-filter microphone arrays is its capability for combating noise sources simultaneously present in the room [1]. This capability stems from the selectivity in spatial volume. This property is corroborated by the results in Fig. 11. Four different values of interfering noise SNR are used, from 20 dB to 0 dB. The interfering noise SNR is calculated over the whole utterance and is defined as the ratio of the power of original speech signal to the power of the original interfering noise signal. That is, the decrease in SNR due to room reverberation is *not* included in this definition. (By this definition, the interference SNR in Fig. 10 corresponds to ∞ dB since there is no competing noise source.) Fig. 11 shows that the correct identification score of the single microphone system decreases as the SNR decreases, while the score of the matched-filter arrays remains almost unchanged. It is also noted from Figs. 10 and 11, that a decrease in SNR from ∞ dB to 20 dB results in a large drop in the identification score for the single microphone system and no change for the matched-filter microphone arrays.[1]

B. Results for Unmatched Conditions

For experiments under unmatched training and testing conditions, close-talking speech is used as the training data. No silence frames are removed when generating the LVQ-based codebooks, since it is hard to remove corresponding silence segments of reverberant speech. The speaker identification

[1]Reverberation (multipath) distortion dominates performance when the competing noise interference is reduced to a low level.

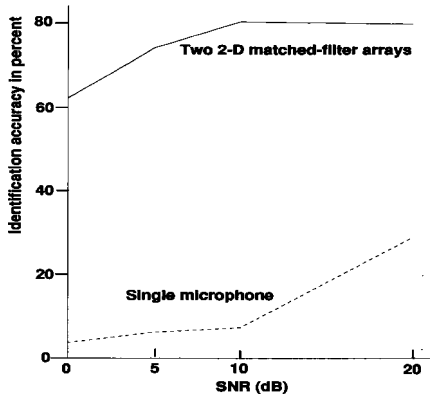


Fig. 12. Speaker identification under an unmatched training and testing condition. A competing noise is presented with varying SNR, see also Fig. 11. No silence removal is made.

results are plotted in Fig. 12. It can be seen that matched-filter arrays, again, perform significantly better than a single microphone. It can also be seen that for SNR $\leq$ 10 dB, the identification result using reverberant speech collected with the single receiver is around the chance level, 1 out of 38 (or 5 out of 190 trails).

However, the overall performance of matched-filter array microphones under an unmatched condition is not good enough to be used practically. It is found that the main factor causing degradation in performance is the silence frames in close-talking speech. If we eliminate these silence frames and then generate reverberant speech with or without a competing noise source, an identification score as high as 90% is noted, under unmatched conditions.

To further improve identification results, we suggest incorporation of a neural network into the system. The neural network can learn and then compensate for multipath distortion. The neural network maps cepstrum coefficients of impaired speech to the corresponding coefficients for close-talking speech, and thereby enhances the performance [16]–[18]. If applicable, one may also retrain the speaker identification system so that it will consistently be used in a matched training and testing condition. Retraining of the speaker identification system is rather simple since only a small amount of speech signal (about 10 s long) is needed.

VI. Conclusion

We have examined as a function of different sound capture systems, the effects of room reverberation and competing noise on the performance of speaker identification. It is found that two 2–D matched-filter microphone arrays, orthogonally placed, are capable of producing high "hands-free" identification scores, even under hostile conditions. Matched-filter array microphones can therefore find use for robust front-end analysis.

We have also studied the effects of vector quantization techniques, codebook size, and order of cepstrum coefficients on speaker identification. It is found that the LVQ is an advantageous vector quantizer for speaker identification, thanks to its flexibility in codebook size. It is also found that the performance of speaker identification depends, in a complex manner, on the size of the codebook and the order of cepstrum coefficients. In the present study, we have not explored use of cepstrum-liftering for reverberant speech. We intend to incorporate bandpass liftering [19] in future work.

The results suggest that the system of microphone arrays and speaker identification can be used favorably in teleconferencing environments as high-quality speaker pickup and for simultaneous speaker identification.

Acknowledgment

The program for the LBG vector quantizer was provided by C. Che. The program for the LVQ quantizer was provided by T. Kohonen and his associates.

References

[1] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, nos. 1–2, pp. 207–222, 1993.
[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE 64*, pp. 460–475, 1976.
[4] A. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE 64*, pp. 475–487, 1976.
[5] S. Furui, "Cepstrum analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 254–272, 1981
[6] G. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE 73*, pp. 1651–1664, 1985.
[7] R. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio, Electroacoust.*, vol. AU–21, pp. 80–89, 1973.
[8] F. Soong, A. Rosenberg, L. Rabiner, and B.-H. Juang, "A vector quantization approach to speaker recognition," in *Proc. of IEEE-ICASSP*, 1985, pp. 387–390.
[9] J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unstrained time-spaced data base," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, 1979, pp. 74–82.
[10] T. Kohonen, J. Kangas, J. Laaksonen, and K. Yorkkola, "LVQ PAK: A program package for the correct application of learning vector quantization algorithms," in *Proc. Int. Joint Conf. Neural Networks*, 1992, pp. 1725–1730.
[11] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantization," *IEEE Trans. Commun.* , vol. COM-28, no. 1, 1980, pp. 84–95.
[12] K. Assaleh and H. Liou (personal communication), 1993.
[13] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.
[14] E. Jan and J. Flanagan, "Image characterization of acoustic multipath in concave enclosures," Rutgers University CAIP Center, Piscataway, NJ, CAIP Tech. Rep. 162, July 1993.
[15] J. Flanagan, D. Berkley, G. Elko, J. West, and M. Sondhi, "Autodirective microphone systems," *Acoustica*, vol. 73, pp. 58–71, 1991.
[16] Q. Lin, E. Jan, C. Che, and J. Flanagan, "Speaker identification in teleconferencing environments using microphone arrays and neural networks," in *Proc. ESCA Workshop Speaker Recognition, Identification, Verification* (Switzerland), Apr. 1994, pp. 235–238.
[17] C. Che, M. Rahim, and J. Flanagan, "Robust speech recognition in a multimedia teleconferencing environment," *J. Acoust. Soc. Am.*, vol. 92, no. 4, pt. 2, p. 2476(A), 1992.
[18] C. Che, Q. Lin, J. Pearson, B. de Vries, and J. Flanagan, "Microphone arrays and neural networks for robust speech recognition," in *Notebook ARPA Human Language Technology (HLT) Workshop* (Princeton, NJ), Mar. 1994, pp. 321–326.
[19] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-35, pp. 947–954.

Qiguang Lin (M'93) was born in Fujian, China, in 1962. He received the B.S. degree from Peking Institute of Technology, Peking, China, in 1982, and the Sc.D. degree in speech communication from Royal Institute of Technology, Sweden, in 1990.

From 1990 to 1992 he was a Research Associate with the Department of Speech Communication and Music Acoustics, Royal Institute of Technology, where he investigated algorithms for articulatory speech synthesis. He is now Research Professor at the Center for Computer Aids for Industrial Productivity, Rutgers University, Piscataway, NJ USA. His current research interests include articulatory modeling of the vocal system, robust speech and speaker recognition using microphone arrays and neural networks, and multimedia communications.

Dr. Lin is a member of Acoustical Society of America and Sigma Xi. His name has been included in Marquis' *Who's Who in the World.*

Ea-Ee Jan (S'93) was born in Taipei, Taiwan, R.O.C., in 1962. He received the B.S. degree in agricultural machinery engineering from National Taiwan University in 1984. He received the M.S. degree in agricultural engineering in 1990 and the M.S. degree in electrical and computer engineering in 1992, both from Rutgers University, New Brunswick, NJ USA, where he is currently working toward the Ph.D. in electrical and computer engineering.

Since April 1992, he has been with the CAIP Center for Rutgers University. His current research interests are in the areas of hands-free sound capture by microphone arrays, speaker recognition, digital signal processing, and parallel computation.

James Flanagan (A'51–M'57–SM'67–F'69–LF'91) received the S.M. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He joined Rutgers University, New Brunswick, NJ USA after extended service in research and research management positions at AT&T Bell Laboratories. He was previously Director of Information Principles Research, with responsibilities in digital communications and information systems. He is Vice-President for Research at Rutgers University and is also Board of Governors Professor in Electrical and Computer Engineering and Director of the CAIP Center. He has specialized in voice communications, computer techniques, and electroacoustic systems, and has authored approximately 150 papers, two books, and holds 45 patents in these fields.

Dr. Flanagan is a Fellow of the Acoustical Society of America and the American Academy of Arts and Sciences. He has received a number of technical awards and is a member of the National Academy of Engineering and the National Academy of Sciences.