

e^{jnt} , where

$$c_n \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} x(t) e^{-jnt} dt$$

for integer n . The Dirichlet kernel [14], $D_n(t)$, is defined by

$$D_n(t) \triangleq \frac{1}{\pi} \sum_{-n}^n e^{jkt} = \frac{1}{\pi} \frac{\sin\left(n + \frac{1}{2}\right)t}{\sin \frac{t}{2}} \quad \text{for } t \in (-\infty, \infty)$$

and all integer n , and it is well known that for integrable $x(t)$ or $[-\pi, \pi]$,

$$S_n(t) \triangleq \sum_{-n}^n c_k e^{jkt} = \frac{1}{2} \int_{-\pi}^{\pi} D_n(t - \lambda) x(\lambda) d\lambda.$$

It follows that if $x(t) = \sum_{-K}^K c_n e^{jnt}$, then

$$x(t) = \frac{1}{2} \int_{-\pi}^{\pi} D_N(t - \lambda) x(\lambda) d\lambda \quad (13)$$

for any $N \geq K$; that is, the kernel $\frac{1}{2} D_N(t)$ for $N \geq K$ reproduces, under convolution on $(-\pi, \pi)$, any 2π -periodic signal $x(t)$ which contains no harmonics beyond the N th. If we take $M_2 = -M_1 = N$ in definition (6) for the reproducing kernel $\phi(t, \tau)$, we find

$$\phi_{-N, N}(t, \tau) = \pi D_N(\omega_0(t - \tau))$$

and the reproducing property (7) is essentially equivalent to (13) since the inner product operation in (7) amounts to a convolution, noting that $\phi(t, \tau)$ depends only on the difference argument $t - \tau$.

Lastly, we remark that the preliminary theorem for harmonic-limited periodic signals, (10), may also be proved as a special case of Kramer's generalized sampling theorem [14]–[15]. This follows since the inner product $\langle x(t), \phi(t, \tau) \rangle$ of (7) defines an integral transform (via Parseval's theorem), namely

$$\frac{1}{T} \int_0^T x(t) \phi^*(t, \tau) d\tau,$$

and we have shown in (8) that $\{\phi(t, nT_1)\}$ is an orthogonal set of elements with respect to our defined inner product, but also orthogonal in the usual sense as well, that is,

$$\frac{1}{T} \int_0^T \phi(t, mT_1) \phi^*(t, nT_1) dt = N\delta_{mn}.$$

The reproducing property of the integral transform kernel is not required for Kramer's result; however, the transition from (10) to the main Proposition relies heavily on this property.

REFERENCES

- [1] H. Stark, "Sampling theorems in polar coordinates," *J. Opt. Soc. Am.*, vol. 69, pp. 1519–1525, Nov. 1979.
- [2] H. H. Stark, J. W. Woods, I. Paul, and R. Hingorani, "Direct Fourier reconstruction in computer tomography," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 237–245, Apr. 1981.
- [3] H. Stark and M. Wengrovitz, "Comments and corrections on the use of polar sampling theorems in CT," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1329–1331, Oct. 1983.

- [4] S. Goldman, *Information Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1953.
- [5] I. T. Turbovitch, "Expansion of periodic functions in a series similar to a Kotelnikov series" (translation), *Telecommun. Radio Eng.*, part 2, vol. 22, pp. 79–82, Aug. 1967.
- [6] N. Maeda, "On sampling theorems of band-limited periodic signals," *J. Inst. Electron. Commun. Eng. Japan*, vol. 50, pp. 1472–1473, 1967.
- [7] J. L. Brown, Jr., "Sampling bandlimited periodic signals—an application of the DFT," *IEEE Trans. Educ.*, vol. E-23, pp. 205–206, Nov. 1980.
- [8] A. Papoulis, *Signal Analysis*. New York: McGraw-Hill, 1977.
- [9] A. Papoulis, "Error analysis in sampling theory," *Proc. IEEE*, vol. 54, pp. 947–955, July 1966.
- [10] E. Hille, "Introduction to general theory of reproducing kernels," *Rocky Mountain J. Math.*, vol. 2, pp. 321–368, 1972.
- [11] H. L. Weinert, Ed., *Reproducing Kernel Hilbert Spaces*. Stroudsburg, PA: Hutchinson & Ross, 1982.
- [12] K. Yao, "Applications of reproducing kernel Hilbert space-band-limited signal models," *Inform. Contr.*, vol. 11, pp. 429–444, 1967.
- [13] L. Baggett and W. Fulks, *Fourier Analysis*. Boulder, CO: Anjou, 1979.
- [14] H. P. Cramer, "A generalized sampling theorem," *J. Math. and Phys.*, vol. 38, pp. 68–72, Apr. 1959.
- [15] A. J. Jerri, "Some applications for Kramer's generalized sampling theorem," *J. Eng. Math.*, vol. 3, pp. 103–105, Apr. 1969.

Speaker-Dependent Isolated Word Recognition Using Speaker-Independent Vector Quantization Codebooks Augmented with Speaker-Specific Data

DAVID K. BURTON AND JOHN E. SHORE

Abstract—A hybrid approach to speaker-dependent isolated word recognition is discussed. The approach merges speaker-specific information obtained from a single training utterance with multisection vector quantization codebooks that were designed for speaker-independent recognition. The approach provides easily trained, computationally efficient, and accurate isolated word recognition. On the digits, the approach achieved an error rate less than 1 percent.

I. INTRODUCTION

For many speakers, recognition error rates less than 1 percent can be achieved for speaker-dependent isolated word recognition of carefully chosen vocabularies [1], [2]. Unfortunately, even on a simple vocabulary such as the digits, error rates this small cannot be achieved when only one or two training utterances are used [3], [4]. Achieving low error rates is usually inconvenient—the speaker must provide several training utterances and the training data often must be processed off line.

Speaker-independent methods are attractive because they eliminate system training by individual users. But there are disadvantages. The error rate for speaker-independent recognition is higher than that for speaker-dependent recognition. Also, speaker-independent systems often require that each unknown input be compared to many reference patterns for each

Manuscript received June 29, 1984; revised September 28, 1984.

The authors are with the Computer Science and Systems Branch, Information Technology Division, Naval Research Laboratory, Washington, DC 20375.

vocabulary word, which means that these systems are computationally expensive and require large amounts of memory.

We have been investigating a hybrid approach that provides easily trained, computationally efficient, and accurate isolated word recognition. Our approach merges speaker-specific information obtained from a single training utterance with multisection vector quantization codebooks that were designed for speaker-independent recognition [5]. Adding speaker-specific data to multisection codebooks is easy because the section codebooks contain unordered sets of spectra that are representative of sounds in the words to be recognized. Using merged codebooks for speaker-dependent recognition of the digits based on a single speaker-dependent training utterance for each digit, we obtained recognition error rates less than 1 percent.

II. BACKGROUND

The approach we take is based on vector quantization (VQ), an information-theoretic data compression principle introduced in the late 1950's [6] and applied recently to speech coding [7], [8] and speech recognition [9], [3], [10], [11], [5]. For an overview of vector quantization, see [12]. Briefly, in VQ, vectors from an information source are represented by a small set of reproduction vectors. The individual reproduction vectors are called *codewords* and the set of reproduction vectors is called a *codebook*. Each codeword consists of linear predictive (LP) parameters that determine the smoothed spectrum of a representative speech frame; the codebook is designed to minimize the average distortion that results from encoding a suitable training sequence [13].

In the multisection VQ approach to isolated word recognition [11], [5], words are recognized by using sequences of VQ codebooks; each word is divided temporally into sections, and each section is represented by a small VQ codebook called a *section codebook*. The sequence of section codebooks is called a *multisection codebook*. A separate multisection codebook is designed for each word in the recognition vocabulary from a training sequence consisting of several repetitions of the word. The resulting multisection codebooks, one for each vocabulary word, are collectively called a *codebook set*. To design a multisection codebook, each training word is first broken into L equal-size and possibly overlapping speech frames, and then it is divided into equal-length sections containing n frames. The first section codebook, for example, is designed from the first n frames of each training utterance, the second codebook from the second n frames. This continues until no training-utterance frames are left. The number of section codebooks is L/n . Based on results from previous studies [5], we used $L = 24$ and $n = 4$.

We used two types of multisection codebooks: clustered and unclustered. In *clustered codebooks*, the codebook design algorithm [13] chooses N codewords that minimize the average distortion for the training data in a particular section where N is specified in advance and $N = 2^r$ for convenience; we call r the section codebook rate. For the distortion measure, we used the gain-normalized Itakura-Saito distortion [14], [3]. *Unclustered codebooks* are generated without the clustering algorithm, simply by making a codeword out of each frame in the training sequence.

After designing a multisection codebook set, unknown words are classified by dividing them into L/n sections of n frames each, encoding them on a section-by-section basis with the multisection codebooks, and finding the multisection codebook that yields the smallest average distortion. For encoding, we used the gain-optimized Itakura-Saito distortion measure [14].

A *merged codebook* consists of the union, on a section-by-section basis, of two or more multisection codebooks. In the work we report here, for each vocabulary word, a speaker-

specific multisection codebook was designed from one training utterance and then merged with a multisection codebook designed from a multiple-speaker training sequence for the same word. We report results using both unclustered and rate-0 (1 codeword) speaker-specific section codebooks. Throughout, LP analysis was based on the autocorrelation method using a Hamming window and the U.S. Navy's 2.4 kbit/s LPC-10 analysis parameters [15]: analysis window width = 16.25 ms, preemphasis factor = 94 percent, and filter order = 10. All speech data were sampled at 8000 samples/s.

The speaker-independent codebooks used in this study were designed using the isolated digits (*zero* through *nine*) in the training portion of a database collected by Texas Instruments (TI) [16]. We tested the merged codebook approach on a separate database, also collected by TI [1], that contains 26 utterances of each of the 10 digits spoken by 16 speakers. We call this the *test* database.

III. EXPERIMENTS

A. Base Line Tests

To establish a base line for comparison, we first performed speaker-independent tests on the test database using codebook sets designed from the training database. We did two experiments: one using rate-3 section codebooks, and the other using rate 5. For each vocabulary word, the training sequence for these speaker-independent codebooks contained two utterances from each of 55 male and 57 female speakers. Using the resulting codebooks, we classified 160 utterances from each of the 16 test speakers. The third and fourth columns of Table I contain the results, which are consistent with previous results [5]. The first 8 speakers are female and the rest are male.

Next we performed two sets of speaker-dependent tests using unclustered codebooks. In the first set, a single utterance of each word was used to build a codebook set for each speaker. In the second, two utterances of each word were used to build the unclustered codebooks for each speaker. The two sets of results are given in columns 5 and 6 of Table I.

B. Merged Codebook Tests

To provide a "good" estimate of the error rates, ten recognition experiments were performed on each speaker; each experiment merged codewords derived from a different set of speaker-specific utterances with a speaker-independent codebook set described in Section III-A. The results we report are averages over these ten experiments.

We first tested the results of merging single-utterance, speaker-specific unclustered codebooks with the rate-3 and -5 speaker-independent codebooks. The results are given in columns 3 and 4 of Table II. In both cases, there was significant improvement over the results from both the speaker-independent and the speaker-dependent approaches (Table I). Note that there is no statistically significant difference between the two merged codebook results. This indicates that small amounts of speaker-specific information can replace much of the information needed for a speaker-independent representation, while at the same time decreasing the error rate. Looking at results for individual speakers, the error rates using merged codebooks are almost always better than both the speaker-independent and speaker-dependent results, and the merged codebook results are always much better when either the speaker-independent or speaker-dependent results are poor. In addition, the merged rate-3 codebooks require only three-eighths the memory and computational complexity of the rate-5 speaker-independent codebooks. The merged codebooks require more memory and distortion computations than the speaker-dependent codebooks, but the merged codebooks achieve significantly smaller error rates.

To see if computational requirements could be further re-

TABLE I
ERROR RATES FOR DIGITS USING SPEAKER-INDEPENDENT AND SPEAKER-DEPENDENT CODEBOOKS

Speaker	No. Class.	Rate 3 Speaker Indep. No. of Errors	Rate 5 Speaker Indep. No. of Errors	1 Utterance Speaker Depen. No. of Errors	2 Utterance Speaker Depen. No. of Errors
ALK	160	3	1	3	0
CJP	160	2	1	12	2
DFG	160	14	7	20	6
GNL	160	0	0	16	1
HNJ	160	14	13	34	9
JWS	160	2	2	3	2
SAS	160	1	1	14	14
SNJ	160	2	0	5	1
GRD	160	22	16	2	1
KAB	160	2	0	1	0
MSW	160	2	5	2	2
REH	160	0	0	3	0
RGL	160	1	1	5	0
RLD	160	2	1	5	1
TBS	160	2	0	1	1
WMF	160	4	6	11	4
All	2560	73 (2.9%)	54 (2.1%)	137 (5.4%)	44 (1.7%)

TABLE II
ERROR RATES FOR DIGITS USING MERGED CODEBOOKS^a

Speaker	No. Class.	Rate 3 + UNC. No. of Errors	Rate 5 + UNC. No. of Errors	Rate 5 + Rate 0 No. of Errors
ALK	160	0.9	0.5	0.9
CJP	160	1.0	1.1	1.0
DFG	160	3.7	1.6	2.8
GNL	160	0.4	0.0	0.0
HNJ	160	7.1	8.1	8.7
JWS	160	1.8	1.1	1.2
SAS	160	0.3	0.2	0.0
SNJ	160	0.2	0.0	0.0
GRD	160	3.9	4.0	7.3
KAB	160	0.1	0.0	0.0
MSW	160	0.2	1.5	2.0
REH	160	0.0	0.0	0.0
RGL	160	0.1	0.1	0.0
RLD	160	0.5	0.0	0.0
TBS	160	0.4	0.1	0.0
WMF	160	2.3	2.2	3.0
All	2560	22.9 (0.9%)	20.5 (0.8%)	26.9 (1.1%)

^aA "+" indicates codebook merging and "UNC." indicates that one-utterance unclustered section codebooks were used. The errors for merged codebooks are averages over ten recognition experiments.

duced, we tried rate-0 clustered instead of unclustered speaker-specific section codebooks. The rate-0 codebooks were made from the same one-utterance training sequences used earlier to make the unclustered codebooks; designing a rate-0 codebook is trivial because the single codeword can be found merely by averaging the autocorrelations of all frames in a section and performing LP analysis on the result [13], [7], [8]. These speaker-specific, rate-0 codebooks were merged with the rate-5 speaker-independent codebooks, and the new merged codebooks were used to classify the same data as before. The results are in the last column of Table II. As expected, they are worse than those using speaker-specific, unclustered codebooks. The results are, however, much better than both the speaker-independent and speaker-dependent results, and for each vocabulary word, only one additional codeword per section and one additional distortion computation per input speech frame are required.

IV. DISCUSSION

Averaged over the 16 speakers, the merged codebook results are significantly better than both the speaker-independent and speaker-dependent results. Also, compared to the rate-5 speaker-independent results, improvement can be achieved with a reduction in memory and computational requirements because smaller speaker-independent codebook rates can be used. In particular, the rate-5 speaker-independent codebooks require 32 distortion computations per input speech frame per vocabulary word, while the rate-3 codebooks merged with unclustered speaker-specific data require only 12. For a typical dynamic-time-warping (DTW) approach [17, Table II], the number of distortion computations per input speech frame is about 10 for each reference template. Thus, the computational complexity of the rate-3 merged codebook approach is about the same as that of a single-reference DTW approach, yet it achieved an error rate less than 1 percent on the digits.

We have described a speaker-dependent isolated word recognition approach that requires little training (one utterance per vocabulary word), achieves a low error rate on the digits (<1 percent), and has small memory and computational requirements. These requirements might be reduced further without a significant increase in error rate by means of the following procedures:

1) reduce the size of merged codebooks by using a clustering procedure [13] on the combined speaker-specific and speaker-independent data,

2) merge a frame (or codeword) of speaker-specific data only when it is significantly different than the existing speaker-independent data.

ACKNOWLEDGMENT

We thank J. Buck for helpful comments, G. Leonard and T. Schalk for help in obtaining the databases, and a referee for his suggestion that the speaker-dependent results be included.

REFERENCES

- [1] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory to practice," *IEEE Spectrum*, vol. 18, pp. 26-32, Sept. 1981.
- [2] W. A. Lea, "Selecting the best speech recognizer for the job," *Speech Technol.*, vol. 1, pp. 10-22, 27-29, Jan./Feb. 1983.
- [3] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 473-491, July 1983.
- [4] A. E. Rosenberg and K. L. Shipley, "Evaluation of an isolated word recognizer in talker-dependent and talker-independent modes using a large telephone band data base," in *Proc. ICASSP 1984, IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, Mar. 1984, pp. 9.5.1-9.5.4, CH1945-5/84/0000-0090.
- [5] D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-word speech recognition using multi-section vector quantization code books," *IEEE Trans. Acoust., Speech, Signal Processing*, 1985, to be published.
- [6] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93-126.
- [7] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 562-574, Oct. 1980.
- [8] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708-721, Nov. 1981.
- [9] A. Buzo, C. Riviera, and H. Martinez, "Discrete utterance recognition based upon source coding techniques," in *Proc. ICASSP 1982, IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 539-542, IEEE 82CH1746-7.
- [10] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol. 62, pp. 1075-1105, Apr. 1983.
- [11] D. K. Burton, J. T. Buck, and J. E. Shore, "Parameter selection for isolated word recognition using vector quantization," in *Proc. ICASSP 1984, IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, Mar. 1984, IEEE 84CH1945-5.
- [12] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, pp. 4-29, Apr. 1984.
- [13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, Jan. 1980.
- [14] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367-376, Aug. 1980.
- [15] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technol.*, vol. 1, pp. 40-49, Apr. 1982.
- [16] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. 1984 ICASSP Conf.*, Mar. 1984, pp. 42.11.1-42.11.4.
- [17] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 623-635, Dec. 1980.

Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator

Y. EPHRAIM AND D. MALAH

Abstract—In this correspondence we derive a short-time spectral amplitude (STSA) estimator for speech signals which minimizes the mean-square error of the log-spectra (i.e., the original STSA and its estimator) and examine it in enhancing noisy speech. This estimator is also compared with the corresponding minimum mean-square error STSA estimator derived previously. It was found that the new estimator is very effective in enhancing the noisy speech, and it significantly improves its quality.

I. INTRODUCTION

Recently [1], we proposed an algorithm for enhancing speech degraded by uncorrelated additive noise when the noisy speech alone is available. This algorithm capitalizes on the major importance of the short-time spectral amplitude (STSA) of the speech signal in its perception, and utilizes a minimum mean-square error (MMSE) STSA estimator for enhancing the noisy speech.

While the distortion measure of mean-square error of the spectra (i.e., the original STSA and its estimator) used in [1] is mathematically tractable, and leads also to good results, it is not the most subjectively meaningful one. It is well known that a distortion measure which is based on the mean-square error of the log-spectra is more suitable for speech processing (e.g., see [2]). Such a distortion measure is therefore extensively used for speech analysis and recognition. For this reason, it is of great interest to examine the STSA estimator which minimizes the mean-square error of the log-spectra in enhancing noisy speech. The derivation of the above STSA estimator and its comparison with the MMSE STSA estimator derived in [1] are the subjects of this paper. This idea of utilizing the above distortion measure for speech enhancement purposes was first proposed in [3] and independently in [4].

The correspondence is organized as follows. In Section II we derived the MMSE log-STSA estimator. The exponential function of the latter estimator is the desired STSA estimator. In Section III we compare by informal listening the performance of the new estimator with that obtained by using the MMSE STSA estimator from [1]. In Section IV we summarize and draw conclusions.

II. DERIVATION OF MMSE LOG-STSA ESTIMATOR

We use here the same formulation of the estimation problem, and the same statistical model, as in [1]. Specifically, the estimation problem of the STSA is formulated as that of estimat-

Manuscript received May 18, 1984; revised August 14, 1984.

D. Malah is with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32 000, Israel.

Y. Ephraim was with the Technion-Israel Institute of Technology. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.