COMPARATIVE STUDY OF DIFFERENT PARAMETERS FOR TEMPORAL DECOMPOSITION BASED SPEECH CODING

S. Ghaemmaghami *M. Deriche* *B. Boashash*

Signal Processing Research Centre
Queensland University of Technology
2 George st, Brisbane, Q 4001, Australia
shahrokh@markov.eese.qut.edu.au m.deriche@qut.edu.au b.boashash@qut.edu.au

ABSTRACT

Temporal decomposition (TD) is an effective technique to compress the spectral information of speech through orthogonalization of the matrix of spectral parameters leading to an efficient rate reduction in speech coding applications. The performance of TD is function of the parameters used. Although "decomposition suitability" of a parameter set is typically defined on the basis of "phonetic relevance" criterion, it can not be directly used in speech coding. Instead, quality evaluation of reconstructed speech is more appropriate. In this paper, we extend our earlier work in this area and attempt to assess several "popular" spectral parameter sets from the viewpoint of decomposition suitability in very low-rate speech coding using parametric, perceptually-based spectral, and energy distance measures.

1. INTRODUCTION

Temporal decomposition (TD) [1] is a method to model the phonemic evolution of speech on the basis of a time sequence of spectral parameters. The phonemic evolution is represented by a number of time-overlapping compact functions, called *target* or *event* functions, which are interpreted as physical representations of speech *acoustic events* [1].

TD uses a matrix of spectral parameters, $\mathbf{Y}$, to extract the corresponding matrix of event functions, $\mathbf{\Phi}$ [1]:

$$\mathbf{Y} = \mathbf{A}\mathbf{\Phi}, \tag{1}$$

where $\mathbf{Y}$ is the pxN matrix of parameters, $\mathbf{\Phi}$ is the mxN matrix of event functions, $\mathbf{A}$ is the pxm matrix of weightings and N, p, and m represent the total number of frames, number of parameters considered, and the number of events extracted from the speech segment, respectively.

To find $\mathbf{\Phi}$ and $\mathbf{A}$, we need to decompose $\mathbf{Y}$ through orthogonalization. Such a decomposition is performed in two stages. First, the locations of event functions are detected using *Singular Value Decomposition* (SVD), and second, the event functions are refined through an iterative algorithm which minimizes the distance (or error) between the estimated and the original parameter sets [1].

Having refined the $\mathbf{\Phi}$ matrix, the matrix of spectral parameters can be approximated using the $\mathbf{A}$ matrix as:

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi}, \tag{2}$$

We have shown earlier [3] that event functions can be approximated by fixed-width (σ) Gaussian functions. With such an approximation, we only need the event locations to place the Gaussian functions. Therefore, the event refinement stage, which is a time-consuming task, can be eliminated and equation (2) changes to:

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Psi}, \tag{3}$$

where $\hat{\mathbf{Y}}$ is the matrix of estimated parameters and $\mathbf{\Psi}$ is the matrix of approximating functions whose (k,n) elements are given:

$$\psi_k(n) = exp(-(n - n_c)^2/2\sigma^2), \tag{4}$$

which are non-zero only in the interval assigned to the segment for which n_c is the central frame index (n : frame index).

Equations (2) and (3) lead to a TD-based very low-rate speech coding system where spectral information is conveyed by matrix $\mathbf{A}$, which is much smaller in size than matrix $\mathbf{Y}$ [1,3].

As evident from the above, the performance of TD relies basically on the temporal characteristics of the spectral parameters used; this performance has been discussed by several researchers [1,2,4]. However, only in [2] that a comparative study is performed, in which 9 different parameter sets have been compared through *phonetic relevance* evaluation and *parametric distance* measurement which can not be directly used in very low-rate speech coding for two major reasons. First, no speech *quality* test has been carried which is of crucial importance in speech coding applications. Second, some spectral parameter sets, such

as *Cepstrum* coefficients, and different combinations of parameter sets (in event *detection* and speech *synthesis*) have not been considered while they have been reported to enhance the performance [4].

In this paper, we discuss the impact of several parameter sets on speech quality, in TD-based speech coding, from aspects which are related to subjective evaluation of reconstructed speech. To do this, we compare 7 parameter sets using three *distance* measures: *parametric*, *spectral*, and *short-time energy*, in TD-based coding with both *original* and *Gaussian* approximated event functions [3].

2. DISTANCE MEASUREMENT

As mentioned earlier, we have considered in this work three distance measures, *parametric* distance (d_p), perceptually-based *spectral* distance (d_s) using a *Bark-scaled filter-bank* [2,5], and *energy* distance (d_e) [6] defined as follows.

$$d_p(n) = [\sum_{i=1}^{p} |par_{1i}(n) - par_{2i}(n)|^2]^{1/2} \qquad (5)$$

where n is the frame index, p is the number of parameters for each frame, and $par_1(n)$ and $par_2(n)$ are the original and the estimated LAR parameters at frame n, respectively. The main advantage of LAR-based parametric distance measure is that it is quite close to the popular *log-likelihood* distance measure [7].

$$d_s(n) = \sum_{k=1}^{15} |P_{1k}(n) - P_{2k}(n)| \qquad (6)$$

where $P_{1k}(n)$ and $P_{2k}(n)$ represent the power of the original and synthesized speech at the k-th filter output of Bark-scaled filter-bank, respectively.

$$d_e(m) = |log[\overline{E}_1(m)] - log[\overline{E}_2(m)]| \qquad (7)$$

where m is the index of the short-time window applied to speech to compute the short-time energy, $\overline{E}_1$ and $\overline{E}_2$ are relative energy (with respect to peak energy of the utterance) of the original and synthesized windowed speech signals, respectively.

The rationale to include *short-time energy* distance metric is to detect temporal distortion which would not be detected by spectral metrics. Such a distortion may affect *durational* information which is important in both *intelligibility* and *naturalness* [8].

Note that we did not use this energy information with other distance measures, based on previous findings [6]. We rather preferred to have each distance function considered independently.

Categorizing Different Distances

Distance measures, as time functions, basically quantify certain differences between reconstructed and original speech signals within short intervals (frames or segments). So, the mean of each distance function gives a *global* measure of distortion associated with the reproduced speech, which can be used as an objective criterion of speech quality in most coding systems.

But, such a global measure, regardless of the type of distance function used, may not conform with the amount of subjective distortion in phonetically-based systems, due to different emphasis on different parts of speech. Therefore, in this work, we categorize the metrics, based on the class of the signals being analyzed, into four measures as mean-distances at: event locations, voiced parts, unvoiced parts, and v/uv (voiced/unvoiced) transitions, defined below.

Assuming $d(n)$ as the distance function (n = frame index), we define d_{event}, d_{voiced}, $d_{unvoiced}$, and d_{vuv} for the abovementioned distances respectively. d_{event} is the mean of distances at location of events, excluding events whose centroids are within ± 15 msec from v/uv transitions. This ensures that d_{event} mostly corresponds to the events located at steady points. d_{event} is then computed from averaging $d(n)$'s at event centroids (three frames for each event):

$$d_{event} = \frac{1}{3N_v} \sum_{k=1}^{N_v} \sum_{i=\mathbf{t}(k)-1}^{\mathbf{t}(k)+1} d(i) \qquad (8)$$

where $\mathbf{t}$ is the vector whose elements are indices of frames associated with selected events and N_v is its size.

d_{voiced} is equal to the mean of distances related to voiced parts, excluding the portions within ± 15 msec around v/uv transitions. Similarly, $d_{unvoiced}$ is obtained from averaging $d(n)$ over unvoiced parts, again excluding v/uv transitions and their neighbors within ± 15 msec.

Finally, d_{vuv} is given as:

$$d_{vuv} = \frac{1}{3N_{vu}} \sum_{k=1}^{N_{vu}} \sum_{i=\mathbf{u}(k)-1}^{\mathbf{u}(k)+1} d(i) \qquad (9)$$

where $\mathbf{u}$ is the vector whose elements are indices of v/uv transitive frames and N_{vu} is its size.

3. PARAMETER SETS

We used 7 different parameter sets: LPC (Linear Predictive filter coefficients), K (reflection coefficients in an LPC model), A (Area parameters in *tube* model), LA (Log-Area), LAR (Log-Area-ratio), Cepstrum, and BF (Band Filter parameters [2]), in event detection. To synthesize speech, we used two parameter sets, LA and LAR as *synthesis* sets, to avoid possible instability [2,7]. The synthesis sets are also used in event refinement as described in section 1, where original events are used (see below).

In total, we conducted 21 experiments in 3 groups (each composed of 7 experiments) for each speech utterance as follows. In the first group, we used the abovementioned sets to detect event functions by TD technique and used LA parameters to refine the events and to synthesize speech. For the second group, we used LAR parameters in event

refinement and speech synthesis. Finally, in the last group, we used the same 7 parameter sets to determine event locations and approximated events by a fixed Gaussian function with $\sigma = 40$ msec. As noted in section 1, with Gaussian approximated events, the refinement stage is eliminated. So, the synthesis set does not affect the TD performance which leads to have only 7 different experiments for this case.

In all experiments, 10 parameters for each frame, except for BF parameters, were computed and used for both event detection and speech synthesis and an LPC model was used to reconstruct the speech signals.

4. EXPERIMENTAL RESULTS

The global mean of the distance functions are shown in Table 1 and Table 2 using original and Gaussian events, respectively. d_p represents the global mean of parametric distance. d_s and d_e, spectral and energy distance measures, are global means normalized with respect to the maximum value energy function across the utterance. Each row shows the results obtained using a given combination of parameters for *detection* and *synthesis*. Also, to make the results easily comparable, distance values of each distance function are ranked with a number (in Bold), where a smaller number represents a lesser distance.

No.	Parameter Sets	d_p	d_s	d_e
1	LA - LA	1.540-**8**	1.447-**10**	.943-**10**
2	LAR - LA	1.068-**9**	1.461-**11**	.941-**9**
3	LPC - LA	1.107-**12**	1.470-**13**	.972-**12**
4	K - LA	1.088-**11**	1.462-**12**	.954-**11**
5	A - LA	1.165-**13**	1.472-**14**	1.022-**13**
6	CEP - LA	1.081-**10**	1.437-**8**	.922-**8**
7	BF - LA	1.396-**14**	1.441-**9**	1.048-**14**
8	LA - LAR	.696-**1**	1.288-**3**	.630-**2**
9	LAR - LAR	.785-**2**	1.296-**4**	.648-**3**
10	LPC - LAR	.840-**6**	1.316-**6**	.652-**5**
11	K - LAR	.828-**5**	1.297-**5**	.649-**4**
12	A - LAR	.990-**7**	1.324-**7**	.753-**7**
13	CEP - LAR	.804-**3**	1.271-**1**	.608-**1**
14	BF - LAR	.827-**4**	1.287-**2**	.656-**6**

Table 1. Parametric, spectral, and energy global distances with original events and different parameter sets.

No.	Parameter Sets	d_p	d_s	d_e
1	LA	1.336-**5**	1.474-**3**	1.995-**3**
2	LAR	1.241-**1**	1.453-**2**	1.983-**2**
3	LPC	1.328-**4**	1.487-**4**	2.004-**5**
4	K	1.262-**2**	1.480-**5**	2.035-**4**
5	A	1.389-**6**	1.499-**6**	2.063-**6**
6	CEP	1.317-**3**	1.479-**1**	1.933-**1**
7	BF	1.577-**7**	1.485-**7**	2.077-**7**

Table 2. Parametric, spectral, and energy global distances using Gaussian events and different parameter sets.

Events	Parameter set	Event centroids	Voiced parts	Unvoiced parts	v/uv changes
Original	LAR	.82	.83	1.25	1.14
	CEP	.79	.88	1.08	1.20
Gaussian	LAR	.79	.79	1.19	1.34
	CEP	.71	.81	1.12	1.41
Mean(p)		.78	.83	1.16	1.27

a. Parametric

Events	Parameter set	Event centroids	Voiced parts	Unvoiced parts	v/uv changes
Original	LAR	1.16	1.46	.28	.51
	CEP	1.19	1.47	.26	.50
Gaussian	LAR	1.26	1.44	.30	.54
	CEP	1.13	1.44	.30	.53
Mean(s)		1.18	1.45	.28	.52

b. Spectral

Events	Parameter set	Event centroids	Voiced parts	Unvoiced parts	v/uv changes
Original	LAR	.73	.48	2.09	1.05
	CEP	.71	.49	2.05	1.09
Gaussian	LAR	.71	.51	1.99	1.08
	CEP	.67	.48	2.00	1.18
Mean(e)		.70	.49	2.03	1.10

c. Energy

Table 3. Normalized modified distances for different classes of speech signals using LAR and Cepstrum parameters as detection sets.

Tables 3a-3c indicate modified distance measures described in section 2, extracted from *parametric*, *spectral*, and *energy* distance functions respectively, for LAR and Cepstrum parameters as event detection sets. The values in these tables are normalized with respect to the global mean of the corresponding distance function. So, a value less than 1 for a particular class of speech shows low sensitivity of the corresponding distance function to that class of signals and vice versa. In other words, the values represent the sensitivity of the distance functions to different classes of signals, leading to better interpretation of the global mean (see section 5).

Figures 1 shows original and Gaussian event functions and reconstructed speech waveforms using Cepstrum coefficients in event detection and LA parameters in event refinement and speech synthesis, for the utterance */she had your dark suit/*.

5. DISCUSSION

The first important observation from Table 1 is the superior performance of LAR parameter set in event refinement

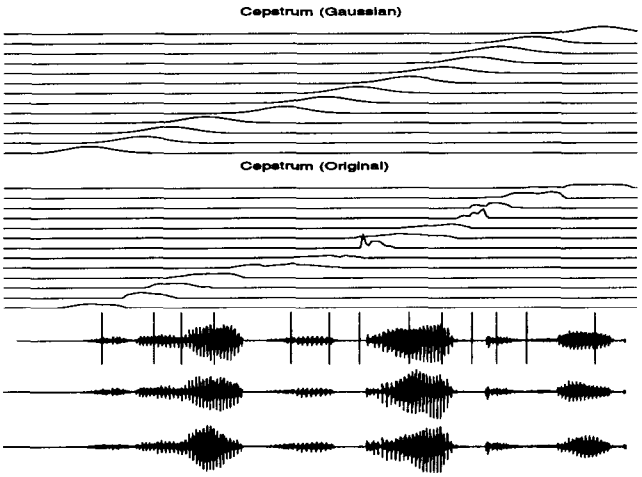

Figure 1. From top: Gaussian approximated and original events, original speech, and reconstructed speech using original and Gaussian events, respectively. Event locations are marked on the original speech waveform.

and speech reconstruction compared to LA's, based on all three distance metrics used in our experiments.

Suitability of LAR in speech synthesis or, in fact, in event refinement, is not equivalent to its performance in event detection. As seen in Table 1, the combination LA-LAR gives the best rank in parametric distance measurement while the best rank in the other two metrics is obtained with the CEP-LAR combination. In other words, the combination of two sets in event detection and event refinement, designates the performance of the method in a certain metric space.

Another important inference from the experiments is in the performance of different metrics. As seen in Tables 3a-3c, parametric distance measure is more sensitive to unvoiced and v/uv changes than voiced and event locations, which means that the global mean of this metric is considerably affected by distortion at unvoiced and v/uv transitions. This clearly shows that such a distance metric does not reflects the "most important" parts of distortion which correspond to event locations and voiced speech as most significant distortions in subjective assessment of intelligibility and naturalness. Indeed, Parametric distance shows the "closeness" of the approximated parameters set to the original set, which is not necessarily equivalent to synthesized speech "goodness" in a phonetically-based coding system, due to non-uniform importance of speech sounds in quality assessment [3,8].

Conversely, spectral distance measure is very sensitive to event locations and voiced parts which shows its suitability in quality assessment in comparison with parametric distance measure in phonetically-based coding. Such a result also resolves the conflict encountered in [2] between the parametric distance measure and *phonetic relevance* test of speech reconstructed using BF and LA parameters as detection sets (compare values of d_p and d_s at rows 1 and 7 in Table 1 where d_s conforms with phonetic relevance test but d_p does not).

Nevertheless, our informal quality assessment shows that even spectral distance measure is not equivalent to subjective quality test in certain conditions. Indeed, certain *temporal* information, which is considered in subjective tests, is lost with such a measure. The problem mostly appears when Gaussian events are used and speech quality degrades due to *palpitation* effect [3]. Such a distortion is better reflected by energy distance measure as indicated in Tables 1 and 2 (compare the ratios of relevant d_e values in two tables to those of d_p d_s). Energy distance measure, although is not suitable to indicate the quality as a unique measure (see Table 3c), it can be used as a complementary to spectral distance measure in some conditions.

6. CONCLUSION

In this paper we have compared 7 different parameter sets as detection sets in TD-based speech coding, from the viewpoint of reconstructed speech quality. We have used three different distance metrics (parametric, perceptually-based spectral, and energy), as objective speech evaluation measures. Based on our experiments, Cepstrum, BF, and LA parameters were found to be the best sets in TD for event detection, while LAR parameters are used in event refinement and as synthesis set in speech reconstruction.

7. REFERENCES

[1] B. S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition", *Proc. ICASSP 83*, pp. 81-84, 1983.

[2] A. M. L. Van Dijk-Kappers, "Comparison of Parameter sets for Temporal Decomposition", *Speech Comm. 8, No 3*, pp. 204-220, 1989.

[3] S. Ghaemmaghami, M. Deriche, "A New Approach to Very Low-Rate Speech Coding Using Temporal Decomposition", *Proc. ICASSP'96*, Vol. 1, pp. 224-227, May 1996.

[4] Y. M. Cheng, D. O'Shaughnessy "Short-Term Temporal Decomposition and its Properties for Speech Compression", *IEEE, Trans. SP*, SP-39, No. 6, pp. 1281-1290, 1991.

[5] A. Sekey, B. A. Hanson, "Improved 1-Bark band-width Auditory Filter", *J. Acoust. soc. Am.*, 75(6), pp. 1902-1904, 1984.

[6] N. Nocerino, F. K. Soong, L. R. Rabiner, D. H. Klatt, Comparative Study of Several Distortion Measures for Speech Recognition", *Proc. ICASSP 85*, pp. 25-28, 1985.

[7] J. R. Deller, Jr., J. G. Proakis, J. H. L. Hansen, "Discrete-Time Processing of Speech Signals", *MacMillan Pub. Co.*, 1993.

[8] D. G. Childers, K. Wu, "Quality of Speech Produced by Analysis-Synthesis", *Speech Comm. 9*, pp. 97-117, 1990.