PERCEPTION OF PHONETIC FEATURES IN SPEECH CODERS FOR MOBILE COMMUNICATIONS

Maurizio COPPERI and Franco PEROSINO

CSELT - Centro Studi e Laboratori Telecomunicazioni -
Via Reiss Romoli 274 - 10148 Torino (Italy)

ABSTRACT

This paper deals with the simulation in real time and the formal subjective evaluation of two low bit-rate speech coders, viz. LPC and RELP, in a mobile satellite system.

The effects of channel impairments, such as multipath fading and shadowing, on intelligibility scores is evaluated by means of the Diagnostic Rhyme Test. The subjective data have been examined to pinpoint the fidelity with which distinctive features and specific phonetic cues are transmitted. Results show that a RELP coder at 9.6 kbit/s, incorporating an error protection scheme, provides a moderately good quality, while the 2.4 kbit/s LPC vocoder yields a quality that is not felt to be commercially acceptable.

1. INTRODUCTION

In our society, mobile communications have become a need for people and a major objective of research. The perception of coded speech under real-world (noisy) transmission conditions is an important aspect of this area, with several implications into the reliability and quality of existing and/or new services (e.g. mobile satellite communications, cellular mobile telephony, etc.) and into the design of efficient and robust speech coding systems.

In this work, two speech digitizers, that is a Residual Excited Linear Predictive (RELP) coder at 7.2/9.6 kbit/s [1], and a Linear Predictive Coder (LPC) at 2.4 kbit/s [2], have been simulated and assessed through the Diagnostic Rhyme Test (DRT) [3]. Useful diagnostic information on specific quality degradations can also be obtained using phonetically constrained sentences [4], instead of rhyming word pairs. Both RELP and LPC algorithms have been used in mobile communications systems [5,6].

Our major objective was the determination of trade-off relationships between speech intelligibility and channel-capacity requirements in mobile satellite systems. In this context, the main constraints to be faced are due to the available bandwidth and transmitter power of satellite and terminals. The bit rate reduction offered by efficient speech digitizers represents an economic incentive in expanding satellite communications, but the attainable subjective quality is of concern if the service should be extended from professional users to the general public.

More specifically, the coders must be capable of providing acceptable quality also in the presence of multipath propagation and inherent signal fading. This degradation, which typically causes burst errors on the transmission link, can be mitigated by the use of error control techniques. The issue of error control was investigated by exploiting four channel models of increasing complexity to choose the optimal method. Results presented throughout this paper have been obtained using the McCullough model [7], which is characterized by 4 independent parameters and can be used to generate sequences that are similar to real-life error sequences. In particular, in this study we consider two examples of bursty channel environments, the former (channel No. 1) typical of land mobile communications for open area into rural, the latter (channel No. 2) including multipath fading. A brief description of the coders is given in section 2. The DRT structure is described in section 3, and the diagnostic scores are discussed in section 4.

2. CODING TECHNIQUES

Both speech compression algorithms and channel models were simulated in real time on an array processor FPS-120 B connected to a VAX 11/785. These coding systems run in half duplex and use a specific audio processing front end with 14-bit A/D D/A converters. The input speech is band-limited to 200-3400 Hz and sampled at 8 kHz. An automatic gain control circuit permits a suitable reduction of the input dynamic range.

The 2.4 kbit/s LPC is based on a 10th order autocorrelation analysis performed every 22.5 ms, an AMDF pitch extractor with median smoother, and a voiced/unvoiced detector driven by the energy ratio between high and low frequency regions.

The 7.2/9.6 kbit/s RELP coder used in this study performs an 8th order autocorrelation analysis over frames of 25 ms in duration, with Hamming windowing of 37.5 ms. After inverse filtering, a 1000 Hz low-frequency portion (baseband) of the residual signal is quantized and transmitted. The regeneration of the full band excitation signal is performed at the receiver using the spectral folding method [8]. The 9.6 kbit/s RELP incorporates an error protection scheme based upon the combination of bit interleaving and bit protection with error correcting codes. The former mechanism is aimed at splitting a long error burst into several shorter bursts (ideally, into isolated errors), thus allowing, through a sort of "divide-and-conquer" strategy, easier protection of the most important parameters in the data frame. The latter mechanism protects the reflection coefficients k(1) through k(4) using four (15,5) BCH codes, and the r.m.s. value of each frame using a (12,4) code. The first code can correct up to 3 errors, whereas the second can correct 1 or 2 errors. Residual samples are left to the channel mercy. Overall, the frame format of the 9.6 kbit/s RELP consists of 190 bits of speech information, 48 bits of error protection and 2 bits for synchronization. Both the LPC vocoder and the 7.2 kbit/s RELP do not exploit error protection.

3. PROCEDURE

A set of four DRT lists was selected for the experiment. Each list contains 116 pairs of English isolated words, read by native American speakers (2 lists read by males, and 2 read by females). These lists were recorded in a quiet environment using an Altec 659A dynamic microphone without a puff screen.

Six different circuit conditions have been examined, combining the three coding bit-rates with two typical channels, as stated in the introduction. Output signals of the processed stimuli have been recorded on analog tapes and then used for the subjective test. Eight listeners took part in the DRT sessions, that were conducted at the Dynastat Inc. (Austin,Texas) in-house speech evaluation facility.

3.1 Structure of the DRT

The DRT of Voiers [3] is based on discrimination between two rhyming monosyllabic words that differ for the initial consonant. The listener's task is simply to indicate which word has been presented. Word pairs are chosen so that initial consonants differ for only one distinctive feature according to the taxonomy shown in Table 1, in which the sign + means positive (present) state of the feature, the sign means negative (absent) state, and the circle means "doesn't apply". Table 2 shows an example of stimulus words used in DRT.

DRT data can be scored in different ways, according to the investigator's interest. In our work, we want to focus not only on the six major features ,i.e. voicing, nasality, sustention, sibilation, graveness and compactness, which are recognized as essential to phonemic distinction for English, but also on scores for the apprehensibility of a given feature, e.g. sustention, in voiced and unvoiced phonemes, or voicing in frictional and nonfrictional phonemes. That is because a finer examination may often pinpoint particular deficiencies of the speech processor. However, the total score is obtained by averaging the six main diagnostic scores.

4. RESULTS AND INTERPRETATIONS

The gross scores of the six critical phonemic features considered in the DRT are plotted in Fig 1. Score differences over subcategories are highlighted in the discussion, while subdivision of the scores according to the voicing state or specific phoneme cues are shown in Table 3.

Noteworthy are the consistent depressions on the voicing, graveness and sustention components for the conditions No. 5 and 6. In fact, these three features separate the RELP coders from the LPC vocoder.

The voicing feature, distinguishes the voiced consonants from their unvoiced counterparts: /b/ from /p/, /d/ from /t/, /v/ from /f/, etc. For the vocoder conditions, there is a small but consistent bias towards the voicing absent state (i.e., voiced is more frequently perceived as unvoiced). This is due to a significant bias towards the friction absent state.

The graveness feature distinguishes /p/ from /t/, /b/ from /d/, /w/ from /r/, /m/ from /n/, etc. The graveness scores are the smallest for almost all conditions, and this wide gap is primarily due to the inherent difficulty in distinguishing the unvoiced consonant pairs /f/-/θ/ and /p/-/t/. The graveness scores exhibit a bias favoring the absent state, especially for the unvoiced and nonplosive sub-categories (see Table 3). This bias is by far larger for female than for male speakers.

We know that degradations on higher frequency components of voice signals affect the graveness and sibilation features most. Therefore, we can conclude that low scores on these features are also due to the inadequacy of the excitation signal fed into the synthesiser. This is more evident in the LPC algorithm, where the excitation signal is modeled in a rigid and poor way. Also in the RELP, however, the frequency components greater than 1 kHz in the excitation are regenerated in a synthetic manner using the baseband, and this approach is not efficient for certain phonemes and for speakers with high frequency energy concentration. This impairment may be mitigated by a better representation of the true full-band residual signal. Indeed, recent algorithms, such as Multi-pulse [9] or Regular-pulse LPC [10] and Vector Excited Coders [11,12,13], aim at improving the subjective quality at low bit-rates by exploiting a perceptually efficient excitation coding method.

The most significant difference between the three pairs of conditions (1-2, 3-4 and 5-6), is given by the sustention feature, which distinguishes the abrupt weak consonants from their sustained counterparts (/p/ from /f/, /b/ from /v/, /t/ from /θ/). The largest drop is observed from RELP to LPC conditions, and in fact sustention suffers the greatest impairment in vocoded speech. We note that for conditions 2, 3 and 4, the unvoiced sustention feature is affected by a bias towards the absent state, i.e., unvoiced sustained consonants become more like stops. For vocoder conditions 5 and 6, there is also a strong bias towards the voiced present state. This bias is primarily a result of pitch and gain coding, which made most voiced stop consonants (e.g. /b/) sound like continuants (e.g. /v/). Improvements on this effect can be obtained with faster frame update for unvoiced speech and better gain quantization.

The nasality feature, which distinguishes /n/ from /d/ and /m/ from /b/, is the best perceived feature for all conditions.

Scores for the feature compactness relate to the compact-diffuse attribute that serves to distinguish /y/ from /w/, /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc. There are no significant differences between the voiced and unvoiced states of the compactness feature.

The sibilation feature, which distinguishes /s/ from /θ/, /ʃ/ from /k/, etc., shows a bias towards the absent state, indicating that strident consonants can be reproduced with mellow cues. This effect is due to deficiencies of the excitation signal, as discussed for the graveness feature.

The maximum degradation in going from the channel No.1 to the channel No.2 is about 5 points for the sustention feature. In particular, comparing the performance of RELP coders, we note that the error protection implemented on the RELP coder at 9.6 kbit/s seems to be more useful to preserve this feature along with sibilation (for the channel No.1) and graveness (for the channel No.2). In fact, large amounts of consonant feature information are carried in the duration and spectral characteristics of adjacent vowels, as well as in the acoustical manifestations of the consonants. Therefore, the error protection of spectral parameters from k(1) to k(4), particularly adequate for vowels, gives benefits also to certain consonants. Of course, loss of information in the upper frequency formants may cause significant degradations. The robustness of nasality for all the conditions, and of voicing for RELP configurations, is clearly evident. Also compactness, which depends on, among other things, the higher second-formant frequencies, appears somewhat robust for all the conditions.

Overall, the DRT scores show the remarkable robustness of the 9.6 kbit/s RELP system, even in case of multipath fading degradation.

The performance of the LPC system is mainly impaired on the voicing, graveness and sustention features, which are generally quite fragile in all vocoders and sensitive to various forms of speech degradation.

5. CONCLUSIONS

We have simulated in real time two speech coding systems at low bit-rates, suitable for mobile satellite communications. We have evaluated their robustness against typical channel degradations using the DRT facility, and got useful information to trade-off between important issues such as power, bandwidth, quality, complexity and delay. It turns out that a 9.6 kbit/s RELP coder is capable of ensuring very good intelligibility, provided that the most important parameters of the side-information be protected with a combination of bit interleaving and error-correcting codes. Short codes must be used. In fact, in addition to being simpler to decode, short codes are more adequate than long ones when the error probability of the channel is large. In particular, a (15,5) BCH code and a (12,4) code have proven to be suitable for our purposes.

Comparing the DRT scores, it results that two subjective categories are gained by the 9.6 kbit/s RELP over the 2.4 kbit/s LPC system. Indeed, the ability to yield fair quality at 2.4 kbit/s using conventional vocoders remains to be seen. Should this happen, however, it could allow an additional reduction of 4 in power and bandwidth.

Recent speech compression algorithms [9-13] provide high quality speech somewhere between 4 and 8 kbit/s, under ideal transmission conditions. Therefore, future problems to be addressed are those associated with their subjective performance in presence of environmental noise, channel errors and multipath fading.

ACKNOWLEDGMENTS

This study has been supported in part by ESA under ESTEC Contract No. 6098/84/NL. The authors would like to thank W. Kriedte, Agency's representative, for his co-operation. Thanks also go to E. Biglieri and G. Albertengo, from Politecnico of Torino, for the simulation of channel models and error-correcting codes.

REFERENCES

[1] M.Copperi et al., "Medium-rate speech coding simulator for mobile satellite systems", Final Report, ESTEC/Contract No.6098/84/NL, Jan.1986

[2] N.Dal Degan and V.Di Lago, "Design and test of a real-time floating point LPC vocoder", Proc. ICASSP, pp. 97-100, Apr. 1983

[3] W.D.Voiers, "Evaluating processed speech using the Diagnostic Rhyme Test", Speech Technology, pp. 30-39, Jan./Feb. 1983

[4] A.Huggins and R.Nickerson, "Speech quality evaluation using 'phoneme specific' sentences", J.A.S.A. Vol.77, pp.1896-1906, May 1985

[5] F.Yato et al., "Performance evaluation of voice coding schemes applicable to INMARSAT standard-B system", IEE 3rd Int. Conf. on Satellite System for Mob. Commun. Navigation, pp. 162-166, June 1983, London

[6] M.McLaughlin, D.Linder and S. Carney, "Design and test of spectrally efficient land mobile communications systems using LPC speech", IEEE Journ. Selected Areas Commun., Vol. 2, pp. 611-620, July 1984

[7] R.H.McCullough, "The binary regenerative channel", B.S.T.J., Vol. 47, pp. 1713-1735, 1968

[8] R.Viswanathan, A.Higgins and W.Russel, "Design of a robust baseband LPC coder for speech transmission over 9.6 kbit/s noisy channels", IEEE Trans. on Commun., Vol. 30, pp. 663-673, Apr. 1982

[9] B.Atal and J.Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", Proc. ICASSP, pp. 614-617, May 1982, Paris

[10] P. Kroon, E. Deprettere and R. Sluyter, "Regular-pulse excitation: a novel approach to effective and efficient multipulse coding of speech", IEEE Trans. ASSP, Vol. 34, pp.1054-1063, Oct. 1986

[11] M.Copperi and D.Sereno, "Vector quantization and perceptual criteria for low-rate coding of speech", Proc. ICASSP, pp.252-255, Mar. 1985, Tampa (FL)

[12] M.Schroeder and B.Atal, "Code-Excited Linear Prediction (CELP): high-quality speech at very low bit rates", Proc. ICASSP, pp. 937-940, Mar. 1985, Tampa (FL)

[13] M.Copperi and D.Sereno, "CELP coding for high quality speech at 8 kbit/s", Proc. ICASSP, pp. 1685-1688, Apr. 1986, Tokyo

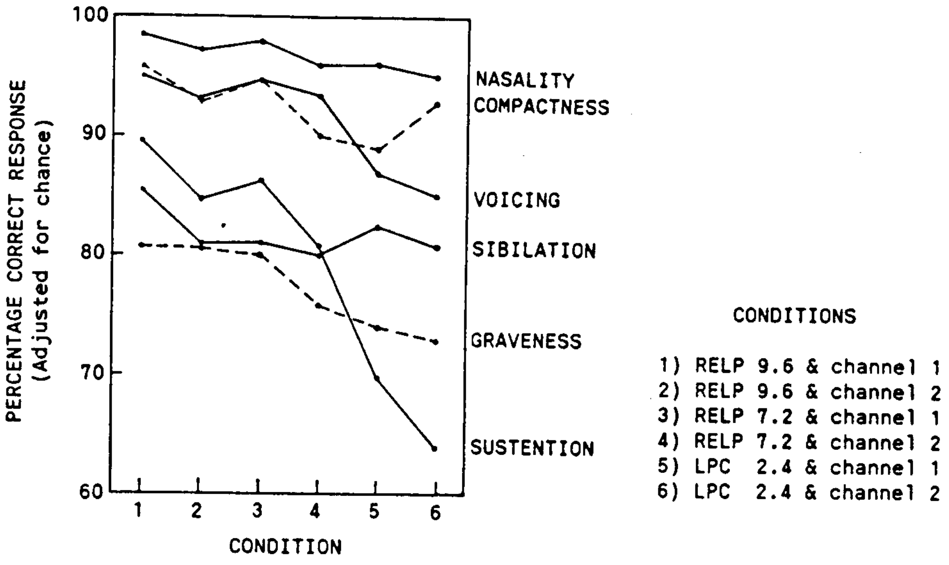

Fig.1 - DRT scores

FEATURES	m	n	v	ð	z	ʒ	ʒ̂	b	d	g	w	r	l	j	f	θ	s	ʃ	ʃ̂	p	t	k	h
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Nasality	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sustention	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+
Sibilation	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Graveness	+	-	+	-	-	0	0	+	-	0	+	-	0	0	+	-	-	0	0	+	-	0	0
Compactness	-	-	-	-	-	+	+	-	-	+	-	-	0	+	-	-	-	+	+	-	-	+	+

Tab. 1 - Consonant taxonomy used in DRT [3]

VOICING voiced-unvoiced	NASALITY nasal-oral	SUSTENTION sustained-interrupted
veal-feel	meat-bear	vee-bee
bean-peen	need-deed	sheet-cheat
gin-chin	nip-dip	vill-bill
dint-tint	moot-boot	thick-tick
zoo-Sue	news-dues	foo-pooh
dune-tune	moan-bone	shoes-choose
goat-coat	neck-deck	those-doze
dense-tense	mad-bad	shaw-chaw
jock-chock	knock-dock	fence-pence
SIBILATION sibil.-unsibil.	**GRAVENESS grave-acute**	**COMPACTNESS compact-diffuse**
cheep-keep	weed-reed	yield-wield
jilt-gilt	peak-teak	key-tea
sing-thing	bid-did	hit-fit
chew-cpo	fin-thin	you-rue
juice-goose	moon-noon	ghost-boast
sole-thole	pool-tool	coop-poop
chair-care	fore-thor	yawl-wall
jab-dab	bond-dong	got-dot
zee-thee	wad-rod	shag-sag

Tab. 2 - Sample of DRT stimulus words [3]

FEATURES	CONDITIONS					
	1	2	3	4	5	6
VOICING	95.1	93.0	94.7	93.4	86.9	85.0
frictional	91.8	89.1	91.0	89.1	82.0	79.7
nonfrictional	98.4	96.9	98.4	97.7	91.8	90.2
NASALITY	98.4	97.3	98.0	95.9	96.1	95.1
grave	98.8	96.9	98.0	94.5	95.7	97.7
acute	98.0	97.7	98.0	97.3	96.5	92.6
SUSTENTION	89.5	84.4	86.3	80.5	69.7	63.9
voiced	86.3	80.1	85.9	77.0	70.3	57.0
unvoiced	92.6	88.7	86.7	84.0	69.1	70.7
SIBILATION	85.5	81.1	81.3	80.1	82.4	80.7
voiced	91.4	88.3	86.7	83.6	85.9	87.9
unvoiced	79.7	73.8	75.8	76.6	78.9	73.4
GRAVENESS	80.5	80.5	80.1	75.8	74.0	72.9
voiced	94.5	96.1	93.0	87.1	87.1	83.2
unvoiced	66.4	64.8	67.2	64.5	60.9	62.5
plosive	84.0	86.3	84.4	80.9	74.2	76.2
nonplosive	77.0	74.6	75.8	70.7	73.8	69.5
COMPACTNESS	95.7	92.8	94.5	90.0	89.1	93.0
voiced	95.7	90.2	95.3	92.6	93.0	96.1
unvoiced	95.7	95.3	93.8	87.5	85.2	89.8
TOTAL SCORE	90.8	88.2	89.2	85.9	83.0	81.7
STD. ERROR	.86	.94	.65	1.03	.72	.77

Tab. 3 - DRT scores of main features and sub-categories