



EVALUATION OF EXPERIMENTAL DIPHONES FOR TEXT-TO-SPEECH SYNTHESIS OF ITALIAN(*)

Pier Luigi SALZA(**), Stefano SANDRI(**) and Enzo FOTI(**)

ABSTRACT

An experiment is described for the performance evaluation of: 1) specifically defined speech units against simple "ideal" diphones for synthesizing vowel to vowel coarticulations and sonorant consonant clusters; 2) "allodiphones" for synthesizing stressed mid vowel allophones /'E/ and /'O/.

By concatenation of properly segmented speech units, 20 test words were synthesized and grouped in 23 pairs, to be evaluated by subjective tests according to a three level paired comparison method. Both "trained" and "untrained" listeners could assign preference to one of the two stimuli of each pair or give no preference.

Results show that in particular contexts triphones provide better fitting of complex coarticulations, while allophones of mid vowels and /r/ require proper "allodiphones", in order to get Italian text-to-speech synthesis of good acoustic quality.

INTRODUCTION

The goal of retailing an exhaustive set of segments for high quality speech synthesis makes rise to two major questions: 1) whether the "ideal" diphone model (ref 1) is suitable for whatever complex sound-to-sound transition, or it must be in certain cases enlarged; 2) how many allophones (and thus proper "allodiphones") must be included in the reference set of a given language.

On one hand, it is well known that in several phonetic contexts, like contiguous vowels and sonorant consonant clusters, coarticulation phenomena are likely to spread beyond the immediately bordering phones, see e. g. (ref 2). On the other hand, it has also been shown that the presence of stress on vowels produces a more or less strong rising of F1 (ref 3), besides an increasing of energy, so that stressed vowels could be considered specific allophones of the language.

From the acoustic point of view the quoted phenomena have been widely analyzed, yet their perceptual effects have been scarcely assessed.

The aim of the present experiment is the performance evaluation by subjective testing of specific speech units against simple diphones, for synthesizing the contexts here described. Two kinds of speech units should be considered:

- a) speech segments larger than the diphone
- b) "allodiphones"

Three parameters must be taken into account for providing the optimal solution:

- I) economy
- II) consistency and coherence of the dictionary
- III) closeness to phonetic reality

(*) Work partially supported by an EEC contract in ESPRIT Project n. 64 "SPIN"

(**) CSELT S.p.a. - Via Reiss Romoli 274, 10148 Torino, Italy

CONTEXT SELECTION

A number of significative cases were elicited from phonotactic occurrences of complex coarticulations, allophonic versions of stressed vowels and liquid consonants in Italian. The selected contexts were embedded in meaningful test words. The speech corpus is listed in Table 1, written with IPA symbols, where the stress mark "' " precedes the stressed vowel.

SPEECH MATERIAL

Different types of speech units for synthesizing the test words were segmented from the center side of nonsense trisyllables, having a fixed structure, beginning and ending with the same vowel, in order to optimize formant stability. The paradigm /VpVp'V/ was chosen as basic reference for segmenting unstressed tokens and /Vp'VpV/ for stressed ones (examples: /ezep'e/ for the diphone /ze/, /ez'epe/ for /z'e/, /opj'olo/ for the triphone /j'ol/, and so on). The nonsense speech samples were embedded in the sentence frame "DEVO DIRE CHIARAMENTE", uttered in a silent room by a professional speaker with declarative intonation, at a constant speaking rate of five syllables per second. The speech material was picked up on high quality video cassettes and then digitally converted at 12 KHz sampling rate. Digital broadband spectrograms and LPC analysis (4 ms frame rate, 15 coefficients) of the speech waveforms were performed, together with computations of intensity and pitch (SIFT algorithm) curves; interactive audiographic facilities were used for segmentation, labelling and storage of the elements in proper LPC format files. Unit boundaries were located by identifying the stable portion in the spectrogram; besides, specific markers were inserted in the LPC representation to identify phoneme and transition boundaries.

By concatenation of proper units, the alternative versions of the 20 test words were generated, resulting in 43 different synthetic stimuli. Segmental duration rules at word level were applied (ref 4 and 5), operating only on the stationary portions of the phonemes. A rightward intrinsic pitch alignment in speech unit concatenation was performed; a pitch adjustment among the stimuli concerning the same word was made, to avoid influences of different pitch contours in similar words.

EXPERIMENTAL SETUP

An experiment was designed for the perceptual evaluation of the acoustic differences caused by the different types of unit concatenation. The 43 synthetic stimuli were grouped in 23 pairs, randomized in 10 replications, 5 times in A-B order and 5 times in B-A order, and automatically recorded on high quality videocassettes. The randomization was made with the constraint that the same word could not be repeated in consecutive stimuli pairs. 11 subjects, both trained and not, listened to the stimuli in a silent room, by means of supra-aural headphone receivers. Each pair was presented to the listeners 5 times (mixed A-B or B-A order); a listening session lasted nearly one hour. 1265 judgments were collected, every pair being evaluated 55 times. Judgments were expressed on a questionnaire after double presentation of first stimulus - second stimulus sequence, assigning preference to one of the two stimuli or giving no preference if the two stimuli were not distinguished.

The listeners were not aware of the particular segments under test, but were asked to evaluate the global acoustic quality of the entire word. The data collected in the experiment were introduced into the computer for the calculation of the preference score.

RESULTS

The preference score of the subjective evaluation is shown in the right side of Table 1. A review of the results, based also on the analysis of spectrograms, is given below according to the context directory reported in the table.

Diphthongs: for economy reasons the test was restricted to the worst case, i.e. stressed contexts were obtained by lengthening of unstressed speech units, according to duration rules, instead of using stressed segments. Thus, when analyzing the results concerning large coarticulations, strong interactions between the concatenation of different speech units and stress effect must be taken into account. Surely, as for rising diphthongs /nw'ov/, /gw'aj/, the triphone was not the preferred solution, also confirmed by low score of the unstressed context /jeg/. Concerning falling diphthongs /'aj/, /r'ej/, some uncertainty comes out, as diphones were never preferred, and a higher preference score is given to triphones in /rej/.

The extremely short duration of semiconsonant plus unstressed vowel coarticulation seems to be the major problem, since "undershoot effect also tends to increase as the temporal proximity of adjacent gesture initiations increases" (ref 6).

Vowel sequences: the high score of "no preference" judgments clearly shows that, in spite of the presence of bordering sonorant consonant, triphones are not necessary for these two-vowel contexts, characterized by greater duration and stronger spectral stability. Sonorant consonant clusters: spectral discontinuities at liquid-vowel boundary in such contexts do not prevent from the use of diphones, which seem largely acceptable. On the contrary, the long sonorant palatal /ɲ:/ is better perceived when triphone is used. Similar behaviour is expected for the liquid cognate /ʎ:/.

Stressed vowels and stressed vowel allophones: the results about these groups of data have to be analyzed together, because two aspects of stress problem are likely to interact strictly: 1) how much the selection of the correct allophonic variants of stressed mid vowels, i.e. /'e/, /'o/, /'ɛ/, /'ɔ/, affects message acceptability; 2) whether stress can be successfully realized by lengthening unstressed tokens or specific stressed segments must be used, being that /'ɛ/ and /'ɔ/ are not produced as unstressed segments.

It comes out that unstressed segments lengthened by rule are more acceptable than the uncorrect stressed "allodiphones", whereas, by comparing stressed tokens, the correct allophone selection is always preferred, and the uncorrect one is always rejected, except when both occurrences can be tolerated by the regional dialect of listeners, as in /ez'ɛmpjo/.

Liquid allophones: in spite of significative acoustic and spectral differences, the several variants of /r/ (namely initial, geminate and intervocalic) require that just the first portion of the phoneme is differentiated, while for the second side of the sound the intervocalic allophone is always acceptable.

CONCLUSIONS

In most of the tested cases, ideal diphones show satisfactory performance, allowing economic solutions together with best benefits. Yet, for some particular contexts the implementation of triphones may guarantee better acceptability of the synthetic message: the falling diphthongs /aj/, /ej/ and the long palatals /ɲ:/, /ʎ:/. As far as stressed vowel allophones are concerned, the best compromise seems to be the use of

stressed speech units for synthesizing /'ɛ/ and /'ɔ/, and unstressed segments for generating all the other stressed vowels. Liquid allophones require just three different diphone types: /#r/, /Xr:/, /Xr/ (X stands for either vowel or consonant). The results of this experiment have been applied for designing a new dictionary of speech units (nearly 1200) for text-to-speech synthesis of Italian.

REFERENCES

1. G E Peterson, W S Wang, E Sivertsen, JASA, 30, 739, (1958).
2. C Fowler, Status Report on Speech Research, SR-81, 1, (1985).
3. M Vayra, 14th Journ'ees d'Etude sur la Parole, GALF, Paris, (1985).
4. P L Salza, 109th Meeting of ASA, Austin, S54, (1985).
5. P L Salza, S Sandri, ICASSP '86, Tokyo, 2035, (1986).
6. B E F Lindblom, M Studdert-Kennedy, JASA, 42, 830, (1967).

Table 1 - Summary of experimental setup.

type of experiment	phonetic contexts (test words)		alternative solutions		% preference score		
			A	B	A	B	no pref
diphones versus triphones or quadriphones	k'auza vor:'ei nw'ovo impjeg'ato gw'aina	1	au + uz	auz	2	38	60
			rē + ēj	rēj	2	71	27
			nw + wō + ov	nwov	7	36	57
			je + eg	jeg	13	11	76
			gw + wa + ai	gwai	51	16	33
	reinv'este bien:'ale	2	re + ei	rei	4	11	85
			ie + en	ien	7	20	73
	af:ret:'arsi s'alpa s'ep:o	3	fr + re	fre	7	11	82
			al + lp	alp	5	36	59
ep: + p:o			ep:o	22	67	11	
allophone verification	ez'empjo maj'olika intr'oiti	4	ze + em	z'e + 'em	49	49	2
			jo + ol	j'ol	89	2	9
			ro + oi	r'o + 'oi	60	9	31
	maj'olika intr'olti ez'εmpjo ap'εrto k'ɔl:a	5	j'ol	j'ɔl	0	100	0
			r'o + 'oi	r'ɔ + 'ɔi	0	95	5
			z'e + 'em	z'ε + 'εm	49	51	0
			p'e + 'er	p'ε + 'εr	9	91	0
	rob'usta kor:'oza k'oro korp'ozo por'ozo	6	#R + ro	#R + Ro	7	13	80
			or: + ro	or: + r:o	5	14	81
			or + Ro	or + ro	5	16	79
			or: + rp	or + rp	25	5	70
			or: + ro	or + ro	4	18	78

R: initial /r/ ; r: intervocalic /r/ ; #: phrase boundary
 1: diphthongs ; 2: vowel sequences ; 3: sonorant consonant clusters ;
 4: stressed vowels ; 5: vowel allophones ; 6: liquid allophones.