SPECTRAL ANALYSIS OF SYNTHETIC SPEECH AND NATURAL SPEECH WITH NOISE OVER THE TELEPHONE LINE

Cristina Delogu ([1]), Andrea Paoloni ([1]), Susanna Ragazzini ([1]), Paola Ridolfi ([2])

[1] Fondazione Ugo Bordoni, v. B. Castiglione 59, 00142 Roma
e-mail: cristina@fub.it

[2] Istituto Superiore Poste e Telecomunicazioni, v.le Europa 190, 00144 Roma

ABSTRACT

In order to explain the different performances obtained with natural and synthetic speech at different linguistic levels over the telephone line, we analyzed the data collected in an experiment where 108 randomized stimuli were presented to 96 subjects. Subjects were required to identify the consonant in 51 CV and 57 VCV meaningful or meaningless words. There were 20 different listening conditions: 6 TTS systems (3 formant-based (SF) and 3 diphone-based (SD)), a pure natural voice (NV) and 3 signal-to-noise (S/N) ratios (6, 0, and -6 dB) for a total of 10 systems, presented both in good and in telephone condition.

The comparison between consonant confusions occurred for natural and synthetic speech with comparable overall levels of intelligibility performance showed that the distributions of the consonant confusions for natural and synthetic speech were often quite different in each condition. Some analyses of different spectrograms suggests that such confusions are due to some problems in the phonetic rules and to the telephone line.

1. INTRODUCTION

In recent years particular attention has been devoted to synthetic speech and much work has been done on evaluating its intelligibility and quality also on the telephone line [1].

One of the foreseen use of TTS technology is for telecommunications network services, which can often require high intelligibility through a telephone channel. Many of these applications (e.g. information systems) require good proper name and address pronunciation when no additional context is given to understand the message [2, 3]. Intelligibility evaluation of synthetic speech over the telephone line is crucial for forecasting the performance of TTS systems in real applications [4].

In order to increase the performance of text-to-speech systems, some more knowledge about the nature of natural and synthetic speech will be helpful. To this purpose, we carried out a consonant confusion test for 19 Italian consonants coarticulated with the vowels /a/, /i/, and /u/, produced by a natural voice with 3 levels of noise and 3 formant-based and 3 diphone-based TTS systems through good and telephone channels.

We use the term diphone-based synthesis for those systems which base sound generation on concatenation of natural speech units. On the other hand, in formant-based synthesis every phoneme and every transition is rule-governed [5]. The main difference between these kinds of synthesis is that diphone-based synthesis has rather high basic quality, since the phoneme transitions are already included in the units themselves, while formant-based synthesis shows much more problems related to our limited knowledge of the whole process of speech production and perception .

Some confusions found in two S/N ratios and in two TTS systems as well as the spectrograms of some signals are discussed in the paper.

2. THE EXPERIMENT

We used an open-response test in which listeners were requested to simply write down what they heard on each trial. In this format all phonemes known to the listener were possible responses.

The experiment was subdivided into three sessions, constituted of 32 trials, 16 in good and 16 in telephone conditions, with each session lasting one hour and allowing the assessment of three systems. Subjects were required to identify the consonant in 51 CV and 57 VCV meaningful and meaningless words. This material allowed to observe confusions among consonants in both initial and medial positions.

2.1 Subjects

96 subjects (48 females and 48 males) of 25 to 45 years of age took part in the experiment. They all were native speakers of Italian and had normal hearing sensitivity bilaterally (an average of 20 dB between 125 and 6000 Hz). Almost all subjects had previous experience in listening to synthetic voices. They all used a computer keyboard for the job. All subjects were employees of an Italian Public Institution who received a day off for their participation.

2.2 Systems

There were 20 different listening conditions: 6 TTS systems (3 formant-based (SF) and 3 diphone-based (SD)), a pure natural

voice, and 3 signal-to-noise (S/N) ratios (6, 0, and -6 dB) for a total of 10 systems, presented both in good and telephone modalities.

2.3 Phonetic material

All systems were tested with 51 CV and 57 VCV meaningful and meaningless words. The CV group contains /p/, /b/, /d/, /t/, /k/, /g/, /tS/, /dZ/, /f/, /v/, /S/, /m/, /n/, /J/, /l/, /L/, /r/ (SAMPA notation). The VCV group contains the same consonant phonemes as in CV plus /s/ and /z/. Both groups had /a/, /i/, and /u/ vowels as environmental context.

All speech files were digitally recorded at our lab with a sampling frequency of 20kHz. The speech files were equalized to approximately 65 dB.

3. CONSONANT CONFUSION

Figure 1 shows the intelligibility decrease in telephone modality for all the systems. In natural speech, the decrease was dependent on S/N ratios, while with the TTS systems the best system (SD1) showed the highest decrease.

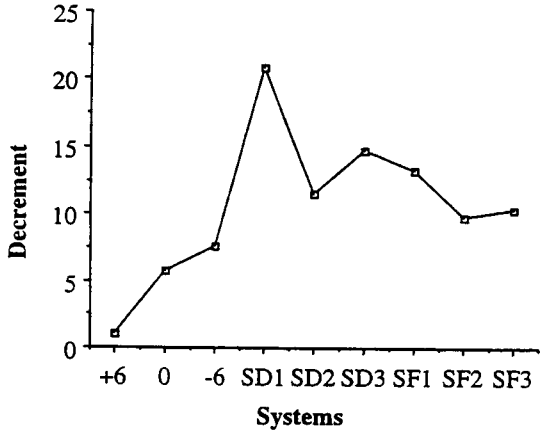


Figure 1: intelligibility decrease from good to telephonic conditions for all systems

The comparison between consonant confusions occurred for natural and synthetic speech with comparable overall levels of intelligibility performance showed that the distributions of the consonant confusions for natural and synthetic speech were often quite different in each condition [6].

Modality	S/N6	SD1	S/N0	SF1
Good	82	92	73	64
Tel	82	76	64	52

Table 1: overall intelligibility scores for the four systems in both modalities.

In order to better analyze such different confusions, we examined the confusion matrices for two S/N ratios (6 and 0 dB) and two TTS systems, SD1 and SF1, since they had comparable overall intelligibility scores, as in Table 1.

The confusion between [ki] and [ti] has been analyzed. Table 2 shows the different incidence of such confusion for the four systems in good and telephone modalities.

Systems	Error Type	Incidence	Error Type	Incidence
Good				
S/N 6	[ki] > [ti]	50%	[ti] > [ki]	0
SD1	[ki] > [ti]	0	[ti] > [ki]	0
S/N 0	[ki] > [ti]	63%	[ti] > [ki]	0
SF1	[ki] > [ti]	0	[ti] > [ki]	94%
Tel				
S/N 6	[ki] > [ti]	13%	[ti] > [ki]	0
SD1	[ki] > [ti]	35%	[ti] > [ki]	0
S/N 0	[ki] > [ti]	44%	[ti] > [ki]	0
SF1	[ki] > [ti]	13%	[ti] > [ki]	44%

Table 2: different incidence in the confusion between [ki] and [ti], for the four systems in good and telephone modalities.

The most interesting finding of those confusions is that in good condition the confusions present in the two S/N ratios are quite different from those in the two TTS systems. The two S/N ratios showed the same confusion a different level of incidence, SD1 didn't show any confusion, and SF1 showed the opposite confusion of that in the S/N ratios. With respect to the telephone condition, the S/N ratios showed the same confusion as in the good condition but with less incidence, SD1 showed the same confusion as that of S/N ratios, and SF1 presented the same confusion as in good condition, but with less incidence.

Furthermore, Table 2 shows some asymmetries in confusions. Asymmetry in consonant confusion means that one sound, e.g. [ki] in the S/N 6 in good condition, is confused with another sound, in our case [ti], while [ti] is never confused with [ki]. This is explained in the literature by the fact that pairs of speech sound sequences which exhibit asymmetries in consonant confusion are acoustically similar except that the ones that are more susceptible to the confusion have one or more distinguishing features which are lacking in the more intelligible sounds [7].

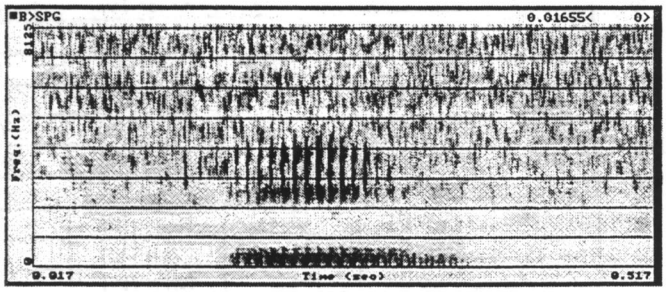
We tried to explain such confusions and asymmetries by analyzing the spectrograms of [ki] and [ti] sounds.

4. SPECTRAL ANALYSIS

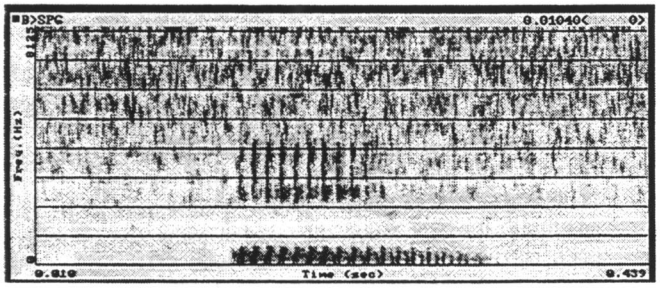
In order to clear up the confusions between [ki] and [ti] the analysis of their spectrograms has been carried out using the CSL speech analysis program. The frame length chosen is 20 ms., the

signal was weighted with a Blackman window and preenphatized with a factor 0.9.

In Figure 2 (a and b) we can compare the spectrograms of natural speech with added noise at S/N equal to 6 dB. We can see that no appreciable difference between the [ki] and the [ti] spectrograms can be noted. This can be due to the fact that noise completely masked that distinct extra feature present in the [k] burst, that in the pure natural voice allow a remarkable distinction between the two phonemes. The similarity between the two spectograms can explain the confusion between [ki] and [ti] showed in the previous paragraph.



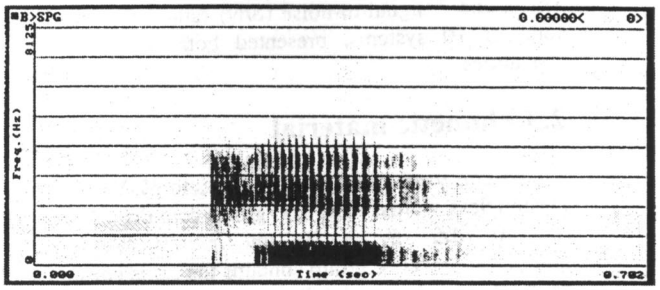
2a)



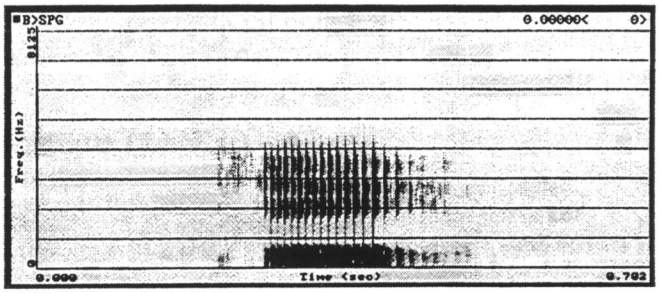
2b)

Figure 2: Spectrograms of natural speech in white noise at S/N 6 dB. In good condition. 2a) [ki], 2b) [ti]

Analyzing the spectrograms of the two syllables produced by SF1, we found that the phoneme [t] has low energy components in a frequency region higher than that in the pure natural speech [t]. This makes the [t] more sounding as a [k] than as a [t], in this case explaining the fact that [ti] was misidentified as [ki].



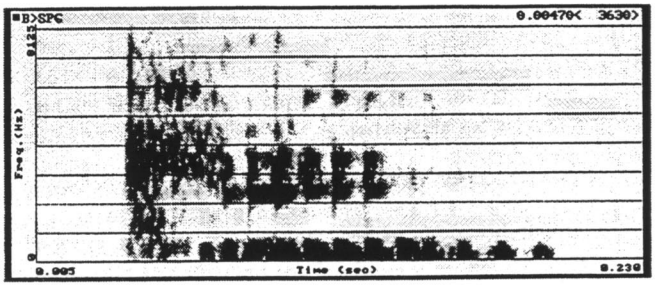
3a)



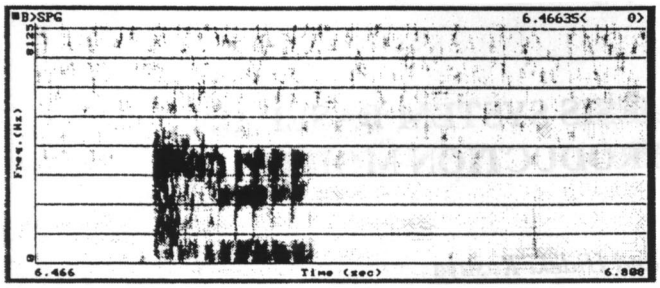
3b)

Figure 3: Spectrograms of synthetic speech produced by SF1 in good condition. 3a) [ki], 3b) [ti]

The telephone channel influences in different manner the intelligibility of all the TTS and natural speech. As in Figure 1, SD1 presented the highest decrease in the telephone line. In Figure 4, where the spectrograms of the good and telephonic realizations of [k] are presented, we can note that the good condition version has a great spectral richness also in high frequency that is completely lost in the telephone version. The intelligibility decrease may be justified in the limitation band of the telephone channel, that reduces the spectral differences between [t] and [t].



4a)



4b)

Figure 4: Spectrograms of synthetic speech [ki] produced by SD1 TT. 4a) good condition , 4b) telephone channel.

5. COMMENTS

In general, intelligibility evaluations give only the overall scores of different systems, while more diagnostic analyses should be useful to the improvement of speech technology. We think in fact that poor intelligibility of synthetic speech means not only that a particular phoneme wasn't recognized, but that it was misidentified as another phoneme. Starting by that, we analyzed the spectrograms of some phonemes that presented confusions both in natural speech with added noise and in two TTS systems.

The analysis of the different spectrograms suggests that the confusions can be due from one hand to the fact that rules used in synthesis don't contain the crucial information that distinguishes two different sound segments and from another hand to the limitation band of telephone channel.

We think that the spectral analysis of sounds that are confused to each other can be useful to better understand and then improve synthetic signals. The work presented here is a work in progress. The spectral analysis of all the confusions occurred for natural and synthetic speech in our consonant confusion experiment are now in processing.

6. REFERENCES

1. Pols L. C. W. "Quality Assessment of Text-to-Speech Synthesis by Rule", In Furui and Sondhi (Eds.) *Advances in Speech Signal Processing*, Dekker, New York, 1992.

2. Spiegel, M.F. and Winslow, E. "Advances in the implementation of effective reverse directory (ACNA) services", Proc. American Voice I/O Society, San Josè, CA, 1995, pp. 145-152.

3. Delogu C., Paoloni A., Ridolfi P., Sementina C. "A Field Evaluation of the Italian 'Automated Reverse Directory Assistance' Service", *International Journal of Speech Technology, in press.*

4. Delogu, C., Paoloni, A., Ridolfi, P. and Vagges, K. "Intelligibility of speech produced by text-to-speech systems in good and telephonic conditions", *Acta Acustica, Vol. 3, No. 1, 1995, pp. 89-96.*

5. Carlson R. "Models of Speech Synthesis". In Roe and Wilpon (Eds.) *Voice Communication between Humans and Machines*, National Academy Press, Washington DC, 1994, pp.116-134.

6. Delogu C., Ridolfi P., Paoloni A. "Confusions among Italian consonants in good and in telephonic conditions: Differences between text-to-speech systems and natural speech with noise". Proc. Eurospeech'95, Madrid, Spain, 1995, pp. 1109-1112.

7. Ohala J. J. "Linguistics and Automatic Processing of Speech" In De Mori and Suen (Eds.) *New systems and Architectures for Automatic Speech Recognition and Synthesis*, NATO ASI Series, Vol. F16 New Systems, Springer-Verlag, Heidelberg pp. 447-475, 1985.