INTELLIGIBILITY OF ENGLISH, FRENCH, GERMAN, AND SPANISH CONSONANTS GENERATED BY RULE OVER SIMULATED TELEPHONE BANDWIDTHS

BATHSHEBA J. MALSHEEN MARISCELA AMADOR-HERNANDEZ MELANIE YUE JAMES T. WRIGHT

Speech Plus, Inc.
640 Clyde Ct.
Mountain View, CA USA 94043

ABSTRACT

The intelligibility of synthetic English, French, German and Spanish initial consonants was tested under normal and telephone bandwidth conditions. Segments were synthesized by language-specific rules using the CallText 5000 text-to-speech converter, based on a cascade-parallel formant synthesizer derived from MITALK-79. The data were compared with those of a human speaker of each language. Nonsense syllables were presented in an open-response format. The results show that (1) on the average synthetic segments for all four languages are 35% less intelligible than human ones, and (2) telephone bandlimiting only slightly degrades synthetic consonants. The findings from nonsense syllables differ from those previously reported for real English words, which were substantially degraded by bandlimiting.

INTRODUCTION

Over the past decade several researchers have tested the segmental intelligibility of a number of English synthesis-by-rule systems [1,2]. The results of these tests have indicated that the intelligibility of high-quality text-to-speech systems is approximately 90% for consonants in real English words. Nevertheless, the level of intelligibility for synthetic phonemes is still significantly lower than that of natural speech. Pisoni, et al suggest that synthetic speech lacks the full compliment of perceptual cues necessary to decode speech. Put in a different way, synthetic speech lacks the acoustic-phonetic redundancy of natural speech, and is acoustically "impoverished".

In order to further investigate the properties of synthetic speech, we measured the intelligibility of synthetic initial consonants for English, French, German, and Spanish, and compared the results of each language to those of a human speaker. In addition, we set out to determine the effects of telephone bandlimiting on the intelligibility of synthetic consonants, in order to learn more about how listeners decode and process these segments. In a previous paper, Wright, et al.[3] found that when consonants in monosyllabic real English words were tested under a telephone bandwidth condition, both human and synthetic consonants suffered significant losses in intelligibility. In this study, we were interested in ascertaining differences in segmental intelligibility between human and synthetic speech when stimuli were nonsense syllables which contain no semantic information to aid in phoneme identification. Finally, we wished to determine whether any consistent phonemic error patterns could be identified for a number of language-specific synthesis-by-rule systems.

Segments were synthesized by rule using the CallText 5000 text-to-speech converter, which is based on a cascade-parallel formant synthesizer derived from MITALK-79 [4]. Whereas the CallText English synthesis-by-rule system is a commercial product which has undergone considerable development and linguistic improvement over the past few years, the other language systems are in earlier stages of development.

METHOD OF TESTING

The intelligibility of initial consonants for all four languages was tested in an open-response format. Test stimuli for each language were CV nonsense syllables which included all of the phonemes occurring in initial position in each language. Three tokens of each phoneme followed by /i,a,u/ were presented to six native speakers in each listening condition. All testing was conducted in a sound-treated room at the Phonology Laboratory of the University of California at Berkeley. None of the subjects had ever before heard synthetic speech.

For English, human and synthetic stimuli were tested in normal and telephone bandwidth conditions. Synthetic stimuli for French, German, and Spanish were tested in normal and telephone bandwidth conditions. In addition, the same stimuli were recorded by French, German, and Spanish native speakers. All recorded human speakers were male.

An average U.S. long-distance telephone line was simulated for the telephone bandwidth condition. Our motivation for simulating a long-distance telephone connection rather than using an actual long-distance line was the variability in transmission performance reported in the "1982/83 End Office Connection Study" [5]. The variability reported suggests that it would be difficult to replicate the actual telephone channel characteristics at different times or to ensure that an actual given connection was typical. The telephone channel introduces many distortions which we potentially could have modelled. We chose to simulate the telephone connection's frequency response and some aspects of its noise characteristics. Our simulator included an octave filter bank to create a transfer function closely matching the average long-distance frequency response reported in the End-Office Connection Study. Because the data in the study represent only the measurements from one central office to the other, the filter bank was adjusted to incorporate the additional losses of the end loops. These losses are estimated to be 1.75 dB per 1000 Hz at each end. The simulator used a codec to maintain a constant signal-to-noise ratio over a wide range of signal levels.

RESULTS

Figure 1 compares mean percentages of error for natural and synthetic consonants in the normal bandwidth listening condition. Note the considerable increase in errors for the synthetic stimuli when compared with the human ones. The differences in mean error percentages between the four synthetic language systems reflect the varying degrees of system development.

Figure 2 compares mean error percentages for English human and synthetic consonants in normal and telephone bandwidth listening conditions. As shown in this figure, error rates for the human stimuli are more affected by bandlimiting than those for the synthetic ones. The mean

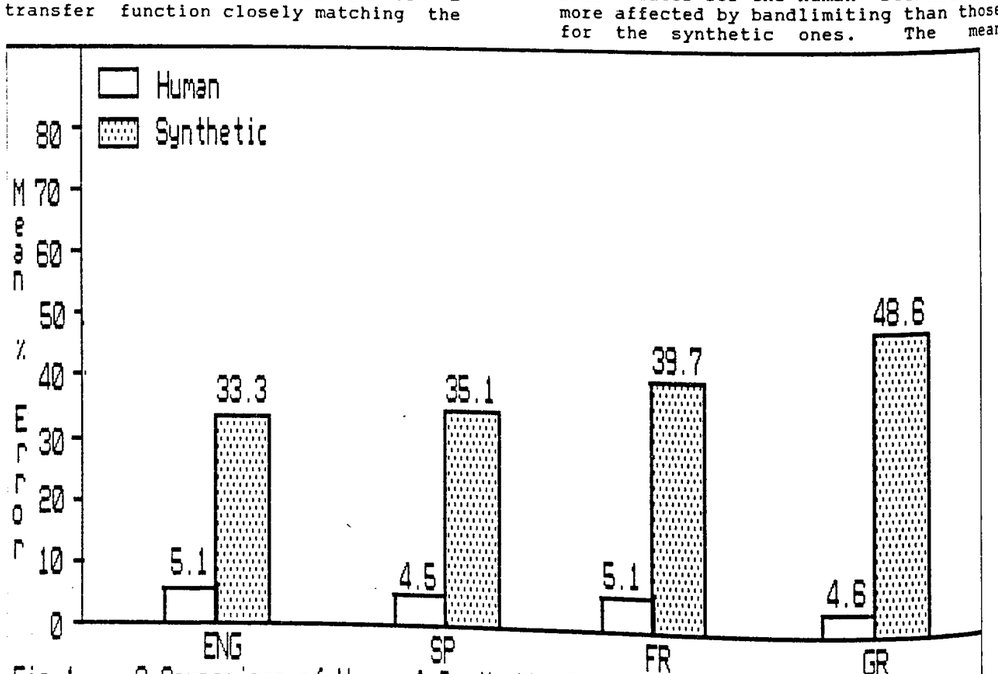

Fig 1 A Comparison of Human & Synthetic Consonants for Four Languages

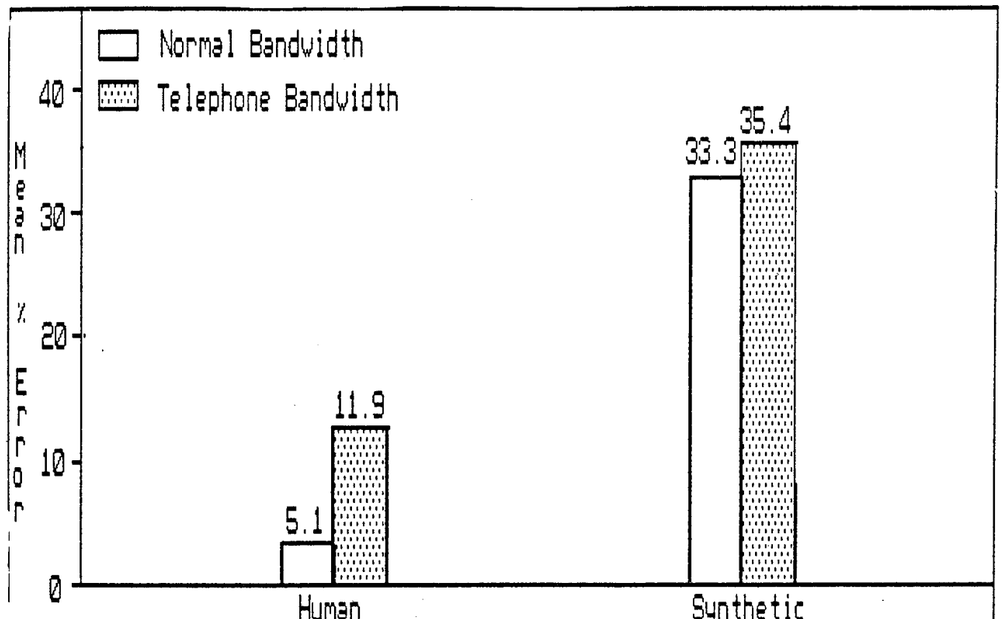

Fig2 Comparison of Eng Human and Synthetic Consonants in Two Bdwth Conds

percentage of error for human consonants increased in the telephone bandwidth condition from 5% to 12%, but the synthetic one increased only from 33% to 35%.

Figure 3 compares intelligibility results for synthetic segments in all four languages for normal and telephone bandwidth conditions. The synthetic segments, which have high error rates in the normal bandwidth condition, degraded only slightly in the bandlimited condition. The average increase in errors for the telephone condition was only 4% from the normal bandwidth condition.

CONCLUSION AND DISCUSSION

Although the data in this study represent only six subjects per language, the results clearly indicate that synthetic nonsense syllables are much less intelligible than human nonsense syllables. Additionally, the data show that the synthetic segments, unlike human ones, suffer only slight degradation due to the attenuation of higher frequencies. These results for synthetic consonants in nonsense syllables are different from those found for real words by Wright, et al. Consonants in real words which were generated by the same synthesizer, degraded substantially when bandlimited. These consonants, however, had intellibility levels of approximately 90% under the normal bandwidth condition. It appears that once intelligibility drops below a certain level the effects of bandlimiting are minimal.

Our findings support Pisoni's contention that synthetic speech is acoustically impoverished. Presumably, a great deal of the acoustic-phonetic information necessary to signal phonemic distinctions is missing or incorrectly specified in the synthetic stimuli. Nevertheless, the fundamental phonetic information which is correctly specified in the synthetic stimuli--the information responsible for the 65% intelligibility of the segments--appears to be sufficiently robust to withstand bandlimiting.

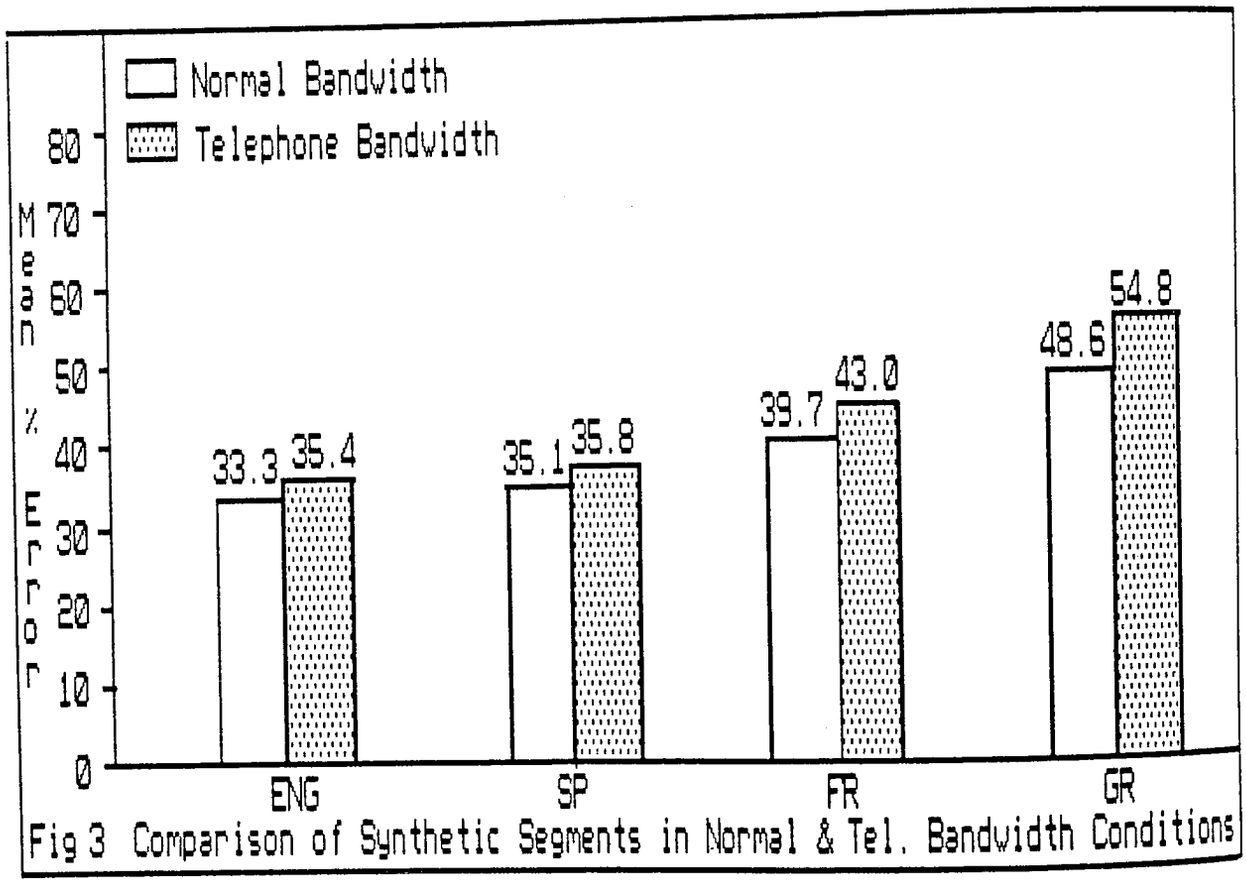

Fig 3 Comparison of Synthetic Segments in Normal & Tel. Bandwidth Conditions

REFERENCES

[1] Pisoni, D.B., Nusbaum, H.C. & Greene, B.G. (1985). Perception of Synthetic Speech by Rule. Proceedings of the IEEE, 73, 1665-1676.

[2] Pisoni, D.B., (1986). Some Measures of Intelligibility and Comprehension. In J. Allen (Ed.), From Text to Speech: The MITALK System, Cambridge U.K.: Cambridge University Press.

[3] Wright, James T., Malsheen, B.J., & Peet, Margot (1986). Comparison of Segmental Intelligibility and Pronunciation Accuracy for Two Commercial Text-to-Speech Systems, Proceedings of American Voice Input/Output Society, 235-261.

[4] Allen, J. (Ed.) (1986). From Text to Speech: The MITALK System. Cambridge UK: Cambridge University Press.

[5] Carey, M.B., Chen, H.J., Descloux, A., Ingle, J.F., & Park, K.I. (1984). 1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network. Bell Systems Technical Journal 63, 2059-2119.