Speech, Hearing and Language: work in progress

Volume 14

EVALUATION OF A MULTILINGUAL SYNTHETIC TALKING FACE AS A COMMUNICATION AID FOR THE HEARING IMPAIRED

Catherine SICILIANO, Geoff WILLIAMS, Jonas BESKOW and Andrew FAULKNER

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

EVALUATION OF A MULTILINGUAL SYNTHETIC TALKING FACE AS A COMMUNICATION AID FOR THE HEARING IMPAIRED

Catherine SICILIANO, Geoff WILLIAMS, Jonas BESKOW and Andrew FAULKNER

Abstract

The goal of the Synface project is to develop a multilingual synthetic talking face to aid the hearing impaired in telephone conversation. This report describes multilingual perceptual studies to characterise the potential gain in intelligibility derived from a synthetic talking head controlled by hand-annotated speech. Speech materials were simple Swedish, English and Dutch sentences typical of those used in speech audiometry. Speech was degraded to simulate in normal-hearing listeners the information losses that arise in severe-to-profound hearing impairment. Degradation was produced by vocoder-like processing using either two or three frequency bands, each excited by noise. 12 native speakers of each of the three languages took part in intelligibility tests in which each of the two degraded auditory signals were presented alone, with the synthetic face, and with a natural video of the face of the original talker.

Intelligibility in the purely auditory conditions was low (7% for the 2-band vocoder and 30% for the 3-band vocoder). The average intelligibility increase for the synthetic face compared to no face was 20%, and was statistically highly reliable. The synthetic face fell short of the advantage of a natural face by an average of 18%. We conclude that the synthetic face in its current form is sufficient to provide important visual speech information.

Introduction

It has been well documented that the visual channel is a uniquely important addition to speech perception. For the hearing impaired community, auditory information is often insufficient for successful communication in the absence of a visual signal. This is particularly relevant for telephone communication, where the hearing impaired user is at a distinct disadvantage. Recent technological developments have shown that the videophone can be a valuable form of communication for hearing impaired people, providing essential visual speech information. However, videophones require expensive equipment at both ends, an impracticality that has led to very limited uptake of the technology. Research has already demonstrated that synthesised visual face movements, driven by an automatic speech recogniser, can be used to deliver phonetic information that is unavailable through the auditory channel to hearing-impaired individuals (Beskow et al., 1997; Agelfors et al., 1998). This technology has the distinct advantage that only the user on the receiving end needs any special technology.

There are two main problems to consider when assessing the potential performance of the synthetic face. A primary concern is the quality and usefulness of the oral movements. We must also consider the accuracy of the automatic speech recogniser driving the facial movements. Even on the simplest of tasks, the most technologically sophisticated large vocabulary continuous speech recognisers, with rich acoustic and language models, will still make errors. For real-time recognition of telephone speech, as is necessary for this application, the task for the speech recogniser is all the more challenging.

Since the synthetic face is driven by the output of the speech recogniser, and this output will certainly contain errors, we should expect that the facial articulations and the speech signal would not be entirely matched. In order to define the potential quality of the synthetic face, however, we first need to examine its usefulness in the ideal case, where the input is error-free. We discuss in this report multilingual perceptual studies exploring the extent of transmission of visual speech information from a synthetic face driven with hand-annotated speech. We examine this property in three languages: British English, Swedish and Dutch. The auditory-visual intelligibility level obtained with the synthetic face is compared with that obtained from the audio signal alone, and with the intelligibility of the audio presented with the natural face of the talker.

The experimental technique consisted of a series of intelligibility tests with native listeners using degraded speech. The use of degraded speech signals with low inherent intelligibility forces the listener to rely more heavily on the visual signal provided by the face; therefore, more information about the utility of the synthetic face is gained under these conditions. The degradation was designed to simulate in normal hearing listeners the reduced intelligibility seen in severe-to-profound hearing impairment by the reduction of spectral detail. Loss of spectral detail is an inevitable consequence of severe-to-profound sensori-neural hearing loss, in which inner hair cell loss causes a broadening of the cochlear filters. This aspect of hearing loss cannot be corrected by amplification. The degradation made use of a vocoder-like technique, with a restricted number of filter bands, and the vocoder always excited by white noise. Such degradation is readily parameterised, and has been extensively used in previous studies, where it has often been considered as a simulation of the information provided by a cochlear implant speech processor (e.g. Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Faulkner, Rosen, & Wilkinson, 2001).

The procedure used here has some advantages over the direct testing of hearing-impaired subjects, since it is possible to present a controlled, consistent signal to each test subject, thus eliminating the variability in individual listeners' hearing abilities that arises when using hearing-impaired subjects.

1. Visual Speech Synthesis

There are two well-documented approaches to visual speech synthesis (see Bailly, 2002 for a review of the state of the art). *Image-based* systems such as VideoRewrite (Bregler et al., 1997) generate video-realistic movies based on images of natural faces, and are the audio-visual analogue of corpus-based speech synthesis. *Model-based* systems on the other hand are based on 3D geometrical models combined with graphical rendering techniques to produce realistic talking heads. A well-known example of this type is Baldi, a descendent of Parke's model, developed at UCSC (Massaro, 1998; Cohen and Massaro, 1993). Systems differ in their approach to the generation of oral movement: these are either biomechanical, in which the underlying musculoskeletal and skin tissue dynamics of speech production are modelled, or purely geometric, with the sole aim of reproducing realistic facial movements without regard to the underlying physiology. The talking head used in this study comes from KTH, Stockholm, and is of the geometrical type. The underlying wire-frame models and smooth-rendered faces are shown in Figure 1. For further details of the implementation of the face, see Beskow (1997).

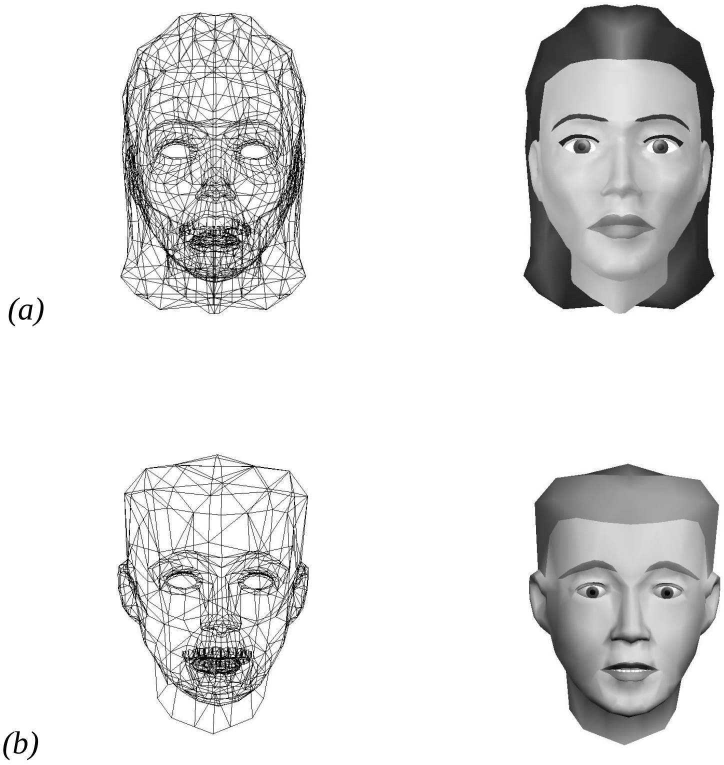

***Figure 1**. Facial synthesis wire-frame and smooth-rendered models (a) Female “Kattis” model (b) Male “Sven” model*

Control software for the face uses articulatory rules to derive oral parameters, which are then mapped to the nodes of the wire-frame face model. For a given language, the target values and time offsets of each of the oral parameters for each phone must be defined in advance. In the current version of the system this is done by hand. The face control software then uses these targets to control the wire-frame face model, interpolating between targets across phones. The software allows for multiple targets within a given phone to be specified for any parameter, thereby capturing some of the within-phone transient properties of speech. In this case, time offsets are specified for each of the targets. Depending on the phone, some of the oral parameters may be left unspecified, thus allowing for the simulation of coarticulatory effects. For example, lip rounding is left unspecified for most consonants. Therefore, lip rounding from a given vowel may “spread” to the surrounding consonants, provided those consonants are not specified for rounding.

To generate an audio-visual sequence, the visual speech synthesis software maps a time-labelled phone sequence to these targets to control the face model, using a simple non-linear interpolation scheme to model coarticulatory effects in a smooth and realistic manner. Visual speech parameters include jaw rotation, labiodental occlusion, lip rounding, bilabial occlusion, tongue tip, tongue length, mouth width, and lip protrusion. A further set of paralinguistic parameters is provided to enable more realistic facial gestures, and include control of brow height, eye position, head angle, and degree of eye closure. These parameters are of secondary interest at this stage in the project and were therefore held constant at neutral levels throughout the studies described here.

2. Method

Seventeen female and nineteen male normal hearing subjects were tested, twelve from each of the language groups. Subjects were all university or institution staff and students, and ranged in age from 18-60. All were native speakers of the relevant language. None of the subjects was familiar with the speech material prior to their participation.

2.1 Speech material

The English test material consisted of the BKB sentence lists (Bench and Bamford, 1979). The sentences were originally designed to measure speech perceptual ability in partially-hearing children, but have also been used quite extensively in speech perceptual tests for normal-hearing listeners. The set consists of 21 lists of 16 sentences, with 50 key words per list (3-4 keywords per sentence). The speaker was a female native speaker of Southern British English.

The Swedish materials were unrelated everyday Swedish sentences developed at KTH by G. Öhngren, based on the principles used in the MacLeod and Summerfield (1990) set for English. The material consists of twelve lists of twelve sentences, with3 keywords per sentence. Additionally, there were 6 lists of 6 sentences used for the practice session. The speaker in this case was a male native speaker of Swedish.

The Plomp and Mimpen sentence lists (Plomp & Mimpen, 1979) were used for the Dutch test material. These sentences are similar in complexity and vocabulary level to the English and Swedish materials. They were designed to test speech perceptual ability in noise. The original material consists of 17 lists of 13 sentences, with 4-7 keywords per sentence. Because portions of the original VHS recording of the materials were corrupted, it was necessary to truncate the material to 15 lists of 12 sentences (including 3 practice). The speaker in this case was a female native speaker of Dutch[1].

2.2 Stimuli

Three separate visual conditions were used: audio-only, synthetic face, and natural face. Each of these was combined with two types of filtered audio, representing different levels of signal degradation. There were thus six test conditions, outlined in Table 1.

The audio-only condition provides a baseline intelligibility level for the degraded speech, while the natural face represents the optimum performance expected to be achievable in practice. The use of two audio conditions allows the performance of the face to be studied over a range of simulated hearing loss. A pilot study comparing 3 and 4-band filters was conducted to determine the appropriate audio conditions. The results indicated the 4-band filter yielded comprehension approaching that found with natural speech, and was thus not a likely to show reliable improvements of intelligibility when combined with visual speech.

[1] These video recordings were made by the Laboratory of Experimental Audiology, University Hospital Utrecht. We are grateful to Guido Smoorenburg for permission to use these recordings in this research.

	Auditory conditions	
Visual conditions	2-band	3-band
Natural face	X	X
Synthetic face from markup	X	X
No face	X	X

Table 1: *Test conditions*

Because the speech material was collected at the relevant institutions for previous projects, the audio sampling rates and video image resolutions used were not uniform. For the purposes of this study, however, these are not important differences.

2.2.1 Audio signal

A noise-excited channel vocoder, implemented in MATLAB 5.1, was used to reconstruct the speech signal as a sum of multiple contiguous channels of white noise over a specified frequency range. Channel characteristics are specified as band-pass filters, the cut-offs and bandwidths of which are set to fit the overall bandwidth and number of channels in terms of basilar membrane distance according to Greenwood's (1990) formula mapping basilar membrane characteristic frequency to basilar membrane position.

The original audio was filtered using 2 and 3 frequency bands, ranging from 100 Hz to 5 kHz, half-wave rectified and with a smoothing filter cut-off frequency of 16 Hz. The filter cut-off frequency of 5 kHz was such that differences across languages in sampling rates for the audio signal were eliminated.

2.2.2 Video signal

The natural face recordings were digitised and compressed with Adobe Premiere using the Indeo 5 codec. The result was a Microsoft AVI file for each of the sentences, with a frame rate of 25 Hz and a resolution of either 360 x 288 (English and Dutch) or 720 x 576 (Swedish). The original audio track was extracted to generate the degraded audio, which was then substituted for the original audio signal of the AVI files.

In order to generate the synthetic face movements from natural speech, time-aligned labels of the audio track were generated for each file. Semi-automatic labelling was performed by means of a forced-alignment procedure using tools from the Speech Filing System software library (http://www.phon.ucl.ac.uk/sfs) and the Wavesurfer package (http://www.speech.kth.se/wavesurfer). The labels were then hand-corrected to yield the time-aligned annotations. The labels are of the same form as would be output by an automatic speech recogniser, assuming the idealisation that the recogniser generates exact time-aligned transcriptions.

These label files were then used to drive the face synthesis, which, for English and Dutch, was captured and stored as an AVI file, and then combined with the degraded audio signal. Frame rate was 25 frames per second, and image resolution was 288 x 360 pixels. Synthesis was performed online for Swedish. For English and Dutch, where the original speaker was female, the Kattis female synthetic face model was used. The original Swedish speaker was male, and therefore the male synthetic face model Sven was used. Both models are driven by the same underlying architecture, so this difference is not expected to have an effect on intelligibility.

3. Procedure

The test subject was seated in front of a computer screen and a loudspeaker or headphones in a soundproofed room or quiet office. The subjects were told that they would hear brief sentences in their native language, and were then asked to repeat what they perceived. The test leader viewed a different screen output on a second monitor, and either entered the subjects' response directly into the computer, or wrote it down. When the answer was noted, the next test sentence was played. The test procedure was controlled by experiment running software developed in Tcl/Tk.

For each language a number of lists was set aside for practice to allow the listeners to accustom themselves to the modified speech signals in each visual condition. In the practice mode, subjects were asked to identify the sentences; for English and Dutch, they were then shown the correct response. This feedback ensured that any learning or adaptation to the conditions would take place before the actual testing began. Practice was given with 2-channel audio, in each of the 3 visual conditions. For Swedish, practice was given for 3-channel audio as well.

Following practice, each subject heard two or three lists in each condition. For English and Swedish subjects , the experiment was run over 2 sessions, with practice given on the first session only. The Dutch experiments were run in a single session. The order of the conditions was randomised for each subject, as was the condition used for each list. The visual conditions—but not the audio—were randomly distributed across the two sessions for the English and Swedish subjects

4. Results

The number of keywords identified correctly was counted using the loose keyword paradigm, whereby errors in morphology are ignored. Scores are expressed as percent of keywords correct out of total possible keywords. The mean results across subjects are given in Table 2 below, grouped by condition and by language. Box-plots of the overall results and for the three languages individually are shown in Figure 2.

The mean number of key words identified by each subject in each condition was entered into a repeated-measures Analysis of Variance (ANOVA), with within-subject factors of auditory signal and visual input, and the between-subject factor of language. This showed highly significant effects of the auditory signal and the visual signal (both at $p<0.001$). The effects of auditory and visual signals each showed significant language group dependencies ($p=0.001$ for auditory signal, $p=0.002$ for the visual signal).[2] Because of this language dependence the effects of auditory and visual signals required to be examined within each language group, by means of 3 separate ANOVAs on the arcsine-transformed intelligibility scores.

[2] These effects remained when mean scores were subjected to an arcsine transformation

	Swedish		English		Dutch	
	2-band	3-band	2-band	3-band	2-band	3-band
audio only	6.0	32.5	14.8	37.2	1.6	19.1
AV: Synface	23.7	60.7	36.7	58.3	15.1	40.2
AV: real face	28.1	66.0	68.4	83.1	32.4	62.5

***Table 2**. Mean percent keywords correct for all subjects*

Within each language group, the effects of auditory and visual signals remained highly significant, and showed no significant interdependence. Planned pairwise comparisons showed that for each language group, the presence of the synthetic face led to a significant increase in intelligibility compared to the absence of a face (always with $p<0.001$). For the Dutch and English groups, the natural face provided significantly higher intelligibility than the synthetic face ($p \alpha 0.001$). For the Swedish group, however, this difference was not significant.

5. Discussion and Conclusions

The data show a significant benefit from the synthetic face under auditory degradation that simulates the speech information available to persons with severe-to-profound hearing losses. Intelligibility on the purely auditory conditions was low (average of 7% for the 2-band vocoder and 30% for the 3-band vocoder) and representative of intelligibility in this target group for the same or similar sentences. The magnitude of the intelligibility increase for the synthetic face compared to no face was broadly consistent, statistically reliable, and of a sufficient magnitude to be important in everyday communication. The average improvement was 20 words out of 100 (range 13.6 to 27.5).

The degree to which the synthetic face fell short of the advantage from the natural face was more variable, with an average of 18 words, but a range of 4.7 to 31.7 words. The variability here is primarily because of the very small differences (of around 5 words) between the synthetic and natural face for the Swedish group. This could be a result of the synthetic face functioning more effectively in the language for which it has primarily been developed, or alternatively because the natural Swedish was relatively difficult to speech-read. The spread of scores for the natural Swedish face in the 3-band vocoder condition is very large compared to that for the natural English face, which does suggest that the Swedish talker is relatively difficult to speech-read, at least for some of the test subjects.

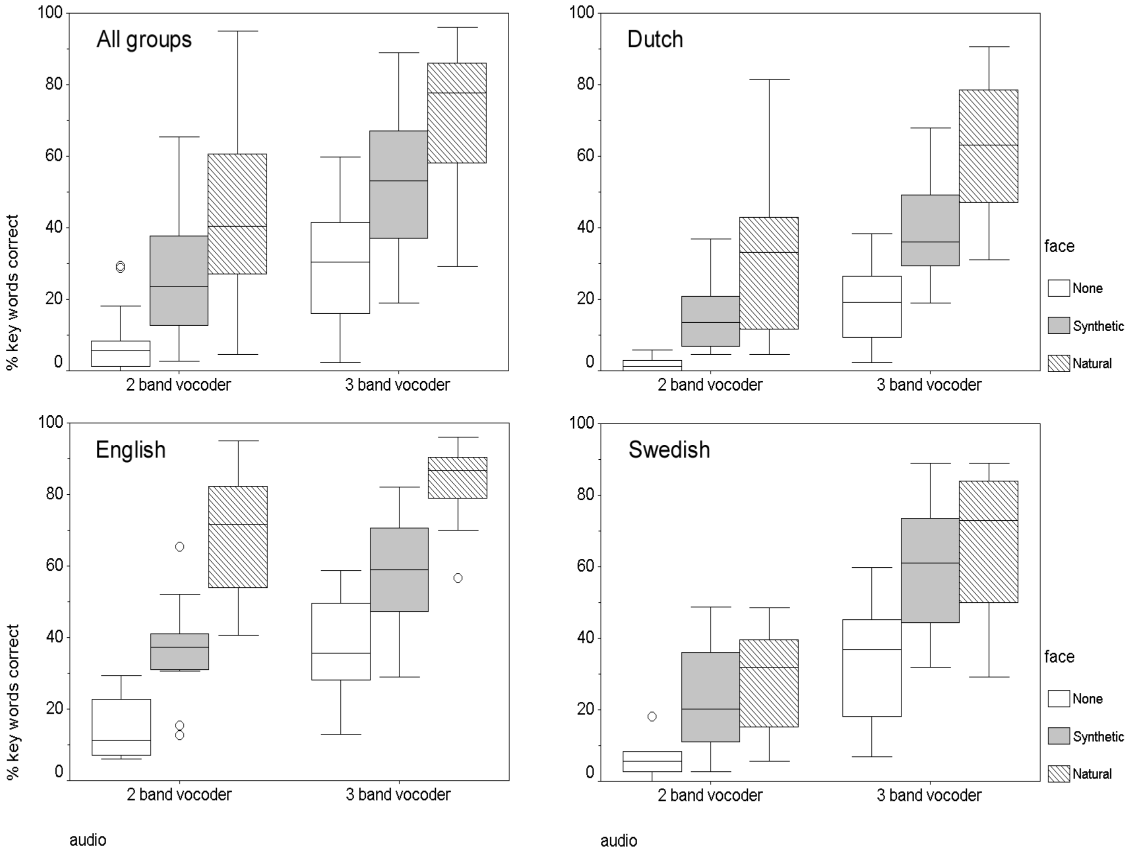

***Figure 2**. Sentence intelligibility for degraded speech with synthetic and natural speech reading cues. The box and whisker plots show the median (bar), interquartile range (box), full range excluding outliers (whiskers) and outlying values (as symbols – outliers are values outside the interquartile range by more than 1.5 times the interquartile range). The upper left panel is the data for all three groups together. The separate language group data are shown in the upper right (Dutch), lower left (English) and lower right (Swedish) panels.*

Since the information transmission from the synthetic face does not approach that from the natural face in the English or Dutch subject groups, it is likely that further refinements of the face synthesis in these languages would be advantageous. At this stage in the project no attempt has been made to optimise the synthesis for these languages, rather a synthesis model developed for Swedish has been somewhat simply extended to include synthesis parameters for segments that do not occur in Swedish. The next phase of activity in this project will cover studies with VCV materials, initially in English, that pinpoint segments for which the synthesis shows shortcomings. The results of these studies will then guide refinements of the visual synthesis.

6. Summary

In summary, we conclude that the synthetic face, controlled by accurate segmental information, is likely in its present form to provide an important communication supplement, both for listeners with profound-to-severe hearing loss and those with normal hearing. Further improvements in the quality of the face synthesis are likely to be possible, and the intelligibility results from the natural face suggest that these will enhance the additional speech information transmitted by the synthetic face.

We are currently extending this work to include hearing-impaired subjects. Preliminary results suggest that those with severe-to-profound hearing impairments are obtaining approximately the same benefit from the synthetic face as the normal hearing subjects in the current study.

Future work will focus on two areas. The first of these is to determine the segments for which the synthesis is poorest, using tests on VCV stimuli and similar materials as mentioned above, so that further development can be focused on these segments. The second involves a systematic examination of the effects of transcription errors on the intelligibility of the synthesised audio-visual speech, with particular emphasis on the types of errors that typically occur in ASR systems.

References

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., Öhman, T. (1998). Synthetic talking faces as lipreading support. *Proceedings of the International Conference on Spoken Language Processing '98.*

Bailly, G. (2002). Audiovisual Speech Synthesis: from ground truth to models. *Proceedings of ICSLP 2002, Denver, Colorado*. pp 1453-1456.

Bregler, C., Cowell, M., and Slaney, M. (1997). VideoRewrite: driving visual speech with audio. *Proceedings of SIGGRAPH '97, Los Angeles CA.* p 353-360.

Bench, J. and Bamford, J. (Eds.) (1979). *Speech-hearing Tests and the Spoken Language of Hearing-Impaired Children.* (Academic, London).

Beskow, J. (1997). Animation of talking agents. In *Proceedings of AVSP '97, ESCA Workshop on Audio-Visual Speech Processing,* Rhodes, Greece.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., Öhman, T. (1997). The Teleface project: Multi-modal speech-communication for the hearing impaired. *Proceedings of Eurospeech '97,* Rhodes, Greece.

Cohen, M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech, in *Models and Techniques in Computer Animatio*n, D. Thalmann and N. Magnenat-Thalmann, (editors). Springer-Verlag: Tokyo. pp.141-155.

Faulkner, A., Rosen, S., & Wilkinson, L. (2001). Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech processor. *Ear & Hearing*, **22**, 431-438.

Greenwood (1990). A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America,* **87**, 2592-2605.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology*, **24**, 29-43.

Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principl*e. Cambridge, MA: MIT Press.

Parke, F. I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics* **2** (9), 61-68.

Plomp, R. and Mimpen, A. (1979). Improving the reliability of testing the speech-reception threshold for sentences. *Audiology*, **18**, 43-52.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science,* **270,** 303-304.