DYNAMICAL SPECTROGRAM, AN AID FOR THE DEAF

A. A. Soltani Farani and E. H. S. Chilton

Centre for Vision, Speech, and Signal Processing
University of Surrey
U.K.

Robin Shirley

Psychology Department
University of Surrey
U.K.

ABSTRACT

Visual perception of speech through spectrogram reading has long been a subject of research, as an aid for the deaf or hearing impaired. Attributing the lack of success in this type of visual aids mainly to the *static* form of information presented by the spectrograms, this paper proposes a system of dynamic visualisation for speech sounds. This system samples a high resolved, auditory-based spectrogram, with a window of 20 milliseconds duration, so that exploiting the periodicity of the input sound, it produces a phase-locked sequence of images. This sequence is then animated at a rate of 50 images per second to produce a movie-like image displaying both the time-varying and time-independent information of the underlying sound.

Results of several preliminary experiments for evaluation of the potential usefulness of the system for the deaf, undertaken by normal-hearing subjects, support the quick learning and persistence of the gestures for small sets of single words and motivate further investigations.

1. INTRODUCTION

Using the eyes, as an alternative to the ears, for speech perception by the deaf has long been a subject of research and development in the electronic age. Two types of visual speech perception aids have been under investigation in recent decades. The first type which is in fact a *sound* visualiser attempts to present sound information in a recognisable form, leaving the task of extracting speech features to the human visual processes. An example of this type is the Visible Speech Translator experiments on which have been reported in the classic book *Visible Speech* by Potter *et al.* [6]. In this technique, the time-frequency information, that is the spectrogram, of an incoming sound is displayed in the form of a conveyor-like image scrolling across a rectangular window and the user follows the sound information by reading this running image. After a considerable amount of research on spectrogram reading, *e.g.* [1, 3, 11], it was found that although this type of speech spectrogram contains sufficient information about the ongoing speech, no one could manage to read the spectrograms with anything like the same effectiveness with which we understand the spoken words.

Two main explanations have been given for the lack of success with spectrogram reading aids. One argument attributed the relatively modest results of the experiments mainly to the poor resolution of the early displays. Another argument fundamentally holds that speech is not a form of simple alphabet to be readable by the eyes, but rather that it is a complex code for which the auditory system serves as a unique decoder [5]. Referring to the precedent of lipreading as a classic example of visual perception of speech, however, it concludes that a more effective type of visual aid might be via a display of the vocal tract articulation.

The second type of visual aid, on the other hand, displays *speech* features automatically extracted from the speech sound in a kind of readable format, for example in the form of cartoon-like animations of the lip, tongue, and jaws *etc.*(lipreading aids [4]), or in the form of written text (speech-to-text translators). The main problem with this type of aid is that the unavoidable task of automatic extraction of speech features is subject to error, particularly in noisy environments.

From these two methods of visual aids, the first method as a sound visualiser has at least two advantages over the second one. The first is that it works independently of the language of the user, unlike a speech-feature display system which is strongly dependent on the language to which it is applied due to the need for robust extraction of the relevant speech features and the phoneme differences between languages. The second is that a sound visualiser system can be used for perception of *sound*, a much more general form of information than speech. This paper, attempts to introduce a new technique of sound visualisation as an aid for speech perception.

2. DYNAMICAL SPECTROGRAM

There are two major problems with the static spectrogram used in the experiments of speech perception aids. The first is that the time resolution of a sliding spectrogram is inevitably limited by its moving speed which is, in turn, determined by the ability of the human eye to follow the running information. To envisage this limit, see Figure 1(a) which shows a duration of 1.9 seconds of the sliding spectrogram of the utterance: ‘one two three’, uttered by a British male speaker. This image scrolls from right to left while the input speech is in progress. Compare the time resolution of this spectrogram, as a maximum practicable resolution, with that of a 60-millisecond segment of it as shown in Figure 1(b) as an auditory perceptible resolution. This inescapable loss of resolution certainly affects the recognition ability of the user while reading the spectrogram [2].

The second and fundamental problem, as we argue, is due to the static, rather than dynamic, nature of the information delivered

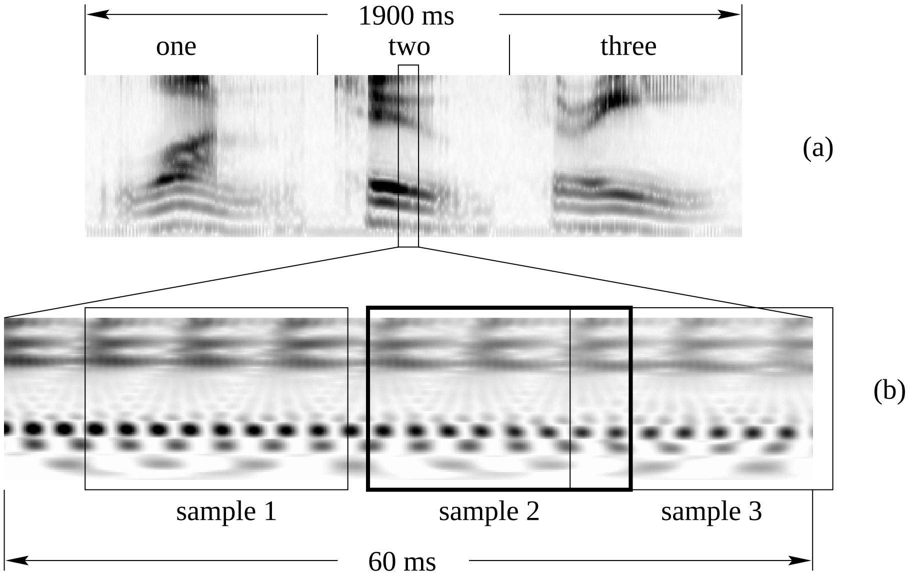

Figure 1: **(a):** Static spectrogram of the utterance: 'One Two Three', spoken by a male speaker. **(b):** A 60-millisecond section of the above utterance in the form of a cochleagram image with a high time resolution. Note the similarity of the three phase-locked samples selected for the animation.

by such a spectrogram. The speech information, by contrast, has primarily a dynamic format which, therefore, has its nature changed if it is illustrated by a sliding *frozen* image such as the static spectrogram. By the same argument, the feasibility of lipreading as an aid for speech perception arises because of its time-varying characteristic [7].

2.1. Auditory Based Dynamical Spectrogram

To solve these two problems we have developed a new system of sound visualisation, the Auditory Based Dynamical Spectrogram (ABDS) [9], through which ongoing speech is translated into a kind of animation which displays the sound information in a properly organised, dynamic format, synchronised with the input sound. In the ABDS system, a sliding spectrogram is first created through a simplified model of the human auditory system, with an arbitrarily high resolution in both the time and frequency domains. This running image, known as the cochleagram [8], is then sampled with a window of 20 milliseconds duration at a rate of 50 samples per second. Each sample is selected from the most recent 40 millisecond of the cochleagram so that the successive images have minimum change between them. Figure 1(b) shows three successive images sampled from a 60-millisecond duration of the cochleagram image. The resulting images are then depicted, one after another, on a screen resulting in an animation which demonstrates both the time-varying and time-independent information of the input sound in real time. By this technique, a steady state sound is translated into a still picture and a spoken word is displayed as a visual gesture.

A slightly modified version of the ABDS has been used for translation of spoken words into visual gestures. This system is called **Audvis** [10]. Figure 2 shows a sequence of 20 images displayed for the word 'one', uttered by a male speaker. The 10 frames in the left column are the representation of the first half of the word, displayed from top to bottom, and the right column displays the second half in the same way.

3. EXPERIMENTS

This system has been evaluated through several experiments, undertaken by normal-hearing subjects, to investigate its potential use as an aid for the deaf or hearing impaired. Although normal hearing subjects are involved, the situation is similar to that of a totally deaf person being tested, because the input sound is not audible to the subjects. In these experiments, recognition of the gestures for a number of words, selected from those likely to be used in the course of seeking directions via a telephone conversation, has been examined through a set of two-word and multi-word forced-choice tests. The experiments have been run through some interactive software, developed for our purposes, capable of managing any number of multi-word tests in three stages, namely, learning, practice, and question [10].

Evaluation Formula

At the question stage of each test a number of questions, T, is given and the number of correct answers selected by the subject, C, is recorded. These numbers as well as the number of forced choices, n, are used to assign a percentage score, P_c, to each test by the formula:

$$P_c = \frac{C - \frac{W}{n-1}}{T} \times 100 \tag{1}$$

where $W = T - C$ is the number of wrong answers. This formula is usually used to compensate for the effect of chance in the evaluation of multi-word, forced-choice tests. By this equation a score is obtained which will be 100 if all the answers are correct, and will have an expected value of zero if all questions are answered randomly.

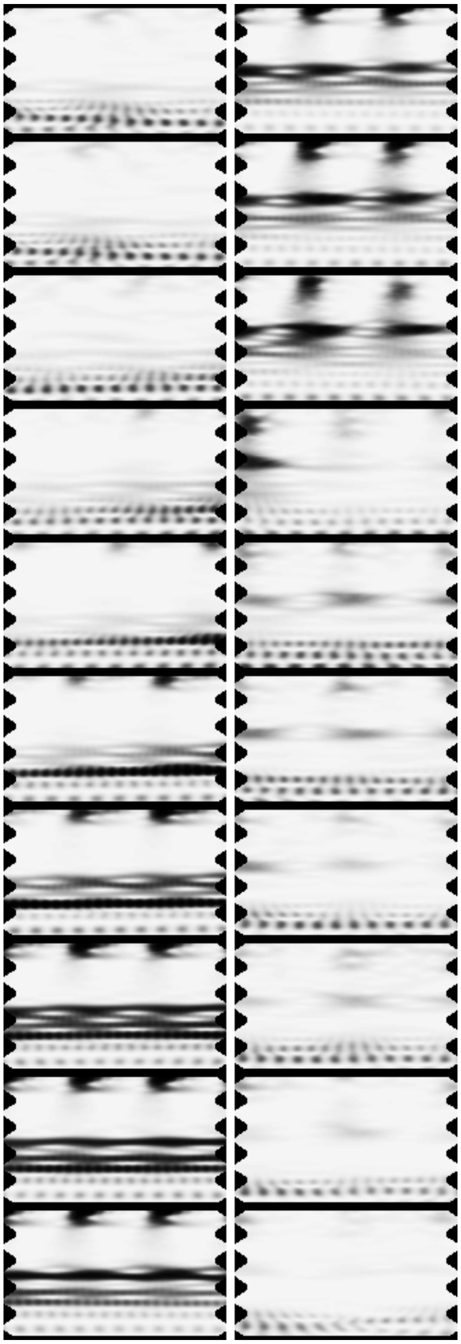

Figure 2: A visual gesture for the word ‘one’, uttered by a male speaker. The 10 frames in the left column are for the first half of the word, displayed from top to bottom, and the right column displays the second half in the same way.

3.1. Experiment 1

This experiment involves 30 novice subjects aged 21-46 years who are totally unfamiliar with the type of visual gestures created by Audvis as translations of corresponding spoken words. It has been aimed at evaluating the ability of normal-hearing subjects, after a certain amount of training, to distinguish between a few words, from any language, displayed as visual gestures. Ten different tests have been given to any subject during a period of about 45 minutes. The words have been selected from four widely different languages: English, Persian, French, and Czech. Each particular test involves a set of two, three, or four known words. Each word of a set is displayed only 10 times in all for learning and 3 times for practice. Then 10-12 random questions, equally distributed over the set of words, are given to the subject. The results are summarised in Table 1.

Words	Mean	SD	repeated trials
yes no	46.7	39.4	the same utterance
yes no	45.3	37.9	different utterances
yes no	49.3	31	different speakers
go stop	80	21.7	different utterances
left right	39.3	29	different utterances
yes no (in Persian)	86.7	21.2	different utterances
yes no (in French)	91.3	22.7	different utterances
yes no (in Czech)	84.7	27.6	different utterances
one two three	68.4	23.3	different utterances
south west north east	57.6	28.6	different utterances

Table 1: Mean and standard deviations of chance adjusted, percentage correct scores for the ten two- to four-word, forced choice tests of Experiment 1. Averages have been computed over 30 subjects. Unspecified words are in English.

For more details about the test procedure and results see [10]. Here, only general conclusions are outlined:

- Average of the recognition scores over 30 subjects were 40-90 percent, where zero percent represents random answers.
- The recognition scores typically increased with the time order of the tests, indicating growth of learning ability as a consequence of more familiarity with the system.
- There was no sign of language dependency of the system as is expected form a sound visualiser system.
- All the words were recognisable after only 10 real-time learning trials, for many of the subjects.
- A negative correlation was observed between the age and the learning ability of users aged more than about 35 years—as was expected.

3.2. Experiment 2

In this experiment, the first author, as a normal-hearing subject sufficiently familiar with the gestures for a number of words, has undertaken a four-word, forced-choice test involving the gestures for four *new* words, namely, 'south', 'west', 'north', and 'east', without any learning and practice trials. By this experiment it was intended to examine the fidelity of the system in translating the auditory cues as correlations between similar words. From 20 random questions given to the subject (selected from among 40 different utterances by the same speaker) only answers to two questions, both involving the word 'south', were wrong (a recognition score of 86.7%). Recognition of these unseen gestures was due to the familiarity of the subject with gestures for such words as: 'left', 'four' and 'three', as they had correlations with the words: 'west', 'north' and 'east', respectively. This experiment and the confusion matrices obtained from the experiment 1 supported the fidelity of the sound visualiser in delivering phoneme information.

3.3. Experiment 3

This experiment has been aimed at evaluation of the persistence of the learned gestures in the long-term memory of the subject. In this experiment, again the first author, well experienced in the recognition of the gestures for the words: 'zero', 'one', . . . , to 'nine', repeated a ten-word forced-choice test, four times during a period of 10 months without any intervening practice during this period. The test involved 20 questions randomly selected from a pool of 100 gestures for the 10 words (10 gestures for each). The results have been summarised in Table 2.

Test no.	Days without practice since previous test	Score (percent)
1	0	100
2	27	94.4
3	58	100
4	86	94.4
5	128	83.3

Table 2: Results from an extended series of repeated recognition tests for the words 'zero' to 'nine', without any practice between tests.

As is evident from the table the gestures have been persistent for durations of less than about four months. Thus it seems that such gestures can persist in the memory for extended periods without refreshing.

4. CONCLUSION

The problem of sound visualisation as an aid for the deaf was addressed. A new technique of Dynamical Spectrogram, as a method of sound visualisation which is based on the primacy of dynamic, rather than static, information in visual perception of speech, was introduced. By this technique a sufficiently resolved, auditory-based spectrogram is created in response to an input sound; then it is sampled with a rectangular window at a rate of 50 samples per second such that a phase-locked sequence of images are produced when a periodic sound is applied to the system. Animation of this sequence on a screen results in a movie-like image which displays a spoken word as a visual gesture and a steady state sound as a still picture. Evaluation of this system through several experiments supported its potential usefulness for visual perception of speech in any language. The visual gestures for a small set of words are recognisable after only 10 real-time learning trials and seem to be persistent in the long-term memory of the user. Finally, the fidelity of the system in translating the phonemes into moving images was confirmed. The results are very promising and motivate further investigations.

Acknowledgement: The first author would like to acknowledge the Ministry of Culture and Higher Education of the Islamic Republic of Iran for their support for this work.

5. REFERENCES

[1] Beth G. Greene, David B. Pisoni, and Thomas D. Carrell. Recognition of speech spectrograms. *J. Acoust. Soc. Am.*, 76(1):32–43, July 1984.

[2] T. Hnath-Chisolm and A. Boothroyd. Speech-reading enhancement by voice fundamental frequency: The effects of f0 contour distortions. *Journal of Speech and Hearing Research*, 35:1160–1168, Oct 1992.

[3] D. H. Klatt and K. N. Stevens. On the automatic recognition of continuous speech, implications from spectrogram-reading experiment. *IEEE transactions on Audio and electro-acoustics*, AU-21(3):210–217, June 1973.

[4] F. Lavagetto. Converting speech into lip movements: a multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering*, 3(1):90–102, March 1995.

[5] A.M. Liberman, F.S. Cooper, D.P. Shankweiler, and Studdert-Kennedy. Why are speech spectrograms hard to read. *Am. Annal Deaf*, 113(2):127–134, 1968.

[6] Ralph K. Potter, George A. Kopp, and Harriet Green Kopp. *Visible Speech*. Dover Publications, Inc., New York, 1966.

[7] L. D. Rosenblum and H. M. Saldaña. Time-varying information for visual speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by Eye II*, chapter 3, pages 61–81. Psychology Press, Sussex, UK, 1998.

[8] M. Slaney and R. F. Lyon. On the importance of time - a temporal representation of sound. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual representations of speech signals*. John Wiley and Sons, New York, 1993.

[9] A. A. Soltani Farani and E. H. S. Chilton. Auditory-based dynamical spectrogram. In *IEEE UK Symposium on Applications of Time-Frequency and Time-Scale Methods (TFTS'97)*, pages 173–176, University of Warwick, UK, 27-29 August 1997.

[10] A. A. Soltani Farani, E. H. S. Chilton, and R. Shirley. Dynamical spectrograms that can be perceived as visual gestures. In *IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (TFTS'98)*, Pittsburgh, Pennsylvania, USA, October 6-9 1998. To be published.

[11] V. W. Zue and R. A. Cole. Experiments on spectrogram reading. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 116–119, 1979.