THE INTELLIGIBILITY OF SHOUTED SPEECH

by

Donald J. MacLean and A. Michael Noll

Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

Xeroxed from The Proceedings of the Symposium on the Aeromedical Aspects of Radio Communication and Flight Safety, AGARD/NATO Advisory Report 19, pp. 10-1 to 10-13, December 1969 (London).

SUMMARY

The results of carefully-controlled laboratory experiments using modified-rhyme intelligibility tests indicate that extremely-shouted speech is less intelligible than normal speech. But while infinite peak clipping degrades the intelligibility of normal speech, the same infinite peak clipping of shouted speech makes it almost as intelligible as normal speech. Large amounts of additive white noise after the peak clipping severely degrade the intelligibility of both the shouted and normal speech almost equally. With shouted speech it is more difficult to identify the speaker. One tempting conclusion to be made from the results of these experiments is that peak clipping of severely shouted speech in a relatively noise-free environment actually is desirable since the clipping improves the intelligibility of the received message. However, during many emergencies the environment is extremely noisy, and peak-clipping of noisy shouted speech might not improve its intelligibility.

BELL TELEPHONE LABORATORIES has often been called upon to decipher and evaluate recorded communications from aircraft that have crashed or caught fire. These recordings have been received from various Federal Agencies. There have been instances of pilots being shot, planes mysteriously out of control, and spacecrafts catching fire. The recordings are usually sent to Bell Telephone Laboratories in an effort to restore the original message from the received garbled communication. These tape recordings are usually of a pilot and/or copilot talking from an aircraft to a ground receiver. Often it is even difficult to understand what the ground receiver is saying even though his channel to the tape recorder is supposedly distortion and noise free. In the case of a pilot shouting over a noisy air-to-ground communication channel in a high background-noise environment, the chances of understanding him are understandably quite small.

Additionally the shouted messages may be even more degraded by inherent nonlinear distortions in the transmitting and receiving equipment. Peak-clipping is often used as a matter of course to increase the long-time average power of the transmitter's modulating system. If the input to the clipping circuit is a flat spectrum or even differentiated the clipping has no pronounced deleterious effect on the message intelligibility.[1] But what happens in an emergency when voice levels are apt to increase in excess of 20 dB? Are the nonlinear distortions of the communication link responsible for the unintelligible message or is it because of an inherently low articulation index for shouting voices or perhaps a combination of both!

Most often only two or three persons at one time, have been involved in the unscrambling work, but occasionally as many as five or six have volunteered their joint cooperation. In these instances all the expertise of years of speech research are brought to bear along with virtually unlimited signal-analysis facilities. Sophisticated computer programs that process the unintelligible messages have been tried but have proved ineffective. These programs have included analysis and synthesis of the speech, time stretching, cepstrum pitch determination, and noise stripping. Additionally, spectrograms have been made of unintelligible portions of the messages in order to try verbal "fits" or educated guesses as to what was originally said. Surprisingly (or perhaps not so surprisingly), the human ear remains the best analyzer although there is a wide disparity between different listeners when the speech is severely garbled. The content of the messages eludes both the human analyzer and sophisticated computers because the original speech was inevitably severely distorted and buried in noise and no amount of known processing can restore the original.

A series of tests has been performed to identify the actual cause of the difficulties encountered in attempting to decipher these emergency messages. More specifically, two series of tests were performed: (1) intelligibility measurements of normal and severely-shouted speech with and without peak clipping, and (2) measurements of the ability of experienced listeners to identify different talkers from samples of both normal and severely-shouted speech. Although some of these tests are still in progress, sufficient data has been obtained and analyzed to report results and draw conclusions; these results and conclusions are reported in this paper with the express hope that a better appreciation of the problems of emergency communications will be conveyed.

INTELLIGIBILITY TESTS

STIMULI AND PROCEDURE. The two main categories of stimuli tested were normal speech and shouted speech. A modified Fairbanks rhyme test was used to test the intelligibility of the consonant-vowel-consonant words.[2] This test consists of six lists of fifty words each, with each list supposedly designed to be equivalent. That is, each list used separately for the same response condition should yield the same intelligibility score. Thus, the subjects have a closed set of six alternatives from which they must select the perceived word. Figure 1 illustrates a typical response sheet used by the listeners. Several versions of this sheet were available with the six words in the boxes randomly permuted to minimize spatial-bias effects.

One speaker shouted the 300 words that comprise this test and also spoke the same 300 words in his normal voice. While he was shouting his auditory feedback or sidetone was deliberately masked by high-level white noise (110 dB SPL) applied via headphones. His instructions were to "shout each word as if he were trying to be heard at the opposite end of a football field." The recordings of the shouted and normal speech were made in the anechoic chamber at the Murray Hill laboratory. Two condenser microphones were used: one spaced 1 meter from the lips of the speaker and the other at 1 centimeter (Bruel & Kjaer type 4131 and type 4135, respectively) as shown in Fig. 2. The microphone at 1 cm closely approximates the distance of the close-speaking microphones usually used in most aircraft communications systems, but it is particularly subject to breath noises. These noises were not particularly noticeable or bothersome for the speaker used in these recordings.

Prior to stimulus presentation, the shouted speech from the close microphone was processed by re-recording the original master tape in the following manner: (1) a

three second interval was inserted between words for subject response time; (2) the peak VU levels of all words were adjusted to be identical; (3) the recording was band-limited from 80 Hz to 10 kHz.

The normal speech was similarly re-recorded except that no gain adjustments were necessary since these words were already very uniform in peak amplitude.

Thirty-six dB of peak clipping (the ratio of the maximum peak-to-peak input to the peak-to-peak clipped output) was used to process both the normal and the shouted speech. In addition, various amounts of white noise were added to produce different signal-to-noise ratios. In the case of clipped-shouted speech, the white noise was added after the clipping.

Four subjects individually took the tests in a quiet office space. They used headphones with the signal level adjusted individually for comfortable listening in their preferred ear (see Fig. 3).

RESULTS. Some of the findings are not surprising in view of the extensive research done on clipped speech. The average percent of words heard correctly by the listeners for normal speech with a residual signal-to-noise ratio of 30 dB was 98.2%, shown graphically in Fig. 4, while normal speech severely clipped averaged 95%. (Pollack & Licklider found infinitely clipped speech to be 92% intelligible.) The two additional curves are the normally-spoken words for two other signal-to-noise ratios: 0 dB and -10 dB. The signal strength in these instances remained constant. For the 0 dB case, the noise level in VU was adjusted to equal the average VU reading of the speech. The -10 dB ratio was set as in the preceding case, but then the noise was simply raised by 10 dB.

Figure 5 illustrates a similar family of curves, only now the results indicate the percent of words heard correctly for shouted speech. Here the intelligibility of the shouted speech for the 30 dB residual noise condition averaged 85.2%; much higher than anticipated. With decreasing signal-to-noise ratios these results similarly follow the normal speech downward in intelligibility. When the shouted speech was peak clipped, the intelligibility increased to 97%, almost as high as normal unclipped speech!

DISCUSSION. One possible reason why clipped shouted speech is more intelligible than unprocessed shouted speech may be that the vowel-to-consonant ratio is restored to more usual levels by the clipping. This can be seen clearly in the spectrograms of Fig. 6. The vowel /i/ in the word heath is shown for normal speech, shouted speech, and clipped shouted speech. This word incidentally was heard incorrectly when shouted but scored correctly when the shouted speech was infinitely clipped.

The spectrogram of the normal speech (spectrogram A in Fig. 6) shows strong formant structures and a rather uniform pitch periodicity (vertical striations). The markings for the initial and final consonants are typical.

The spectrogram of the same word shouted (spectrogram B in Fig. 6) is very different. The formant locations are poorly defined, and the fundamental pitch has risen considerably. Since there are fewer harmonics in the resonant regions (formant bands), the resolution has suffered. The general appearance is more noise-like. Also, the final consonant energy is missing except for a narrow vertical bar at the very end of the utterance. Notice that the duration of the shouted word is twice that of the normally spoken one.

If the shouted word (heath) is clipped, it appears as in spectrogram C. The overall appearance is similar to the previous case, but even more "noisy" in character. This is due in part to the decreased signal-to-noise ratio imposed by the infinite clipping process. Another noticeable difference is that the initial and final consonants are more pronounced; in effect, the consonant-to-vowel ratio has been improved. There also seems to be an improvement in the definition of the formants although not a substantial one.

TALKER RECOGNITION TESTS

STIMULI AND PROCEDURE. To verify if different shouters were recognizable from one another, an additional test was devised and performed. Five volunteers each shouted ten words, namely, BEACH, SIP, TAN, LOT, RAW, HOOK, MOOD, CUT, BIRD, and FILE. They also spoke the same list in their normal voice. These people were recorded in the free-space room using the equipment and conditions described earlier. For this test, however, only the recordings from the microphone 1 meter from their lips were used. The ten words used are similar to some of those in the Fairbanks test and were chosen for their different vowel content and a variety of initial and final consonants. Only ten words were used, since the extreme shouting was physiologically too difficult for prolonged recording sessions.

The fourteen listeners were all colleagues of the talkers and therefore all familiar with the normal speech of the talkers. Nevertheless, the listeners first heard the normal speech over a pair of head phones. During this conditioning phase of the test, the recorded talkers each announced who they were and recited the phrase "Joe took father's shoe bench out - she was waiting on my lawn." This allowed the listener an opportunity to become accustomed to the voices over headphones and also enabled them to adjust the listening level to a comfortable setting.

During the next portion of the test the subjects were asked to identify a particular talker (normal voice) from a randomly ordered set of the 50 stimuli: 5 talkers each speaking 10 words. Approxiamtely 3 seconds was allowed for their choices.

Finally, another 50 word response set was presented to the listeners - this time however the same words were shouted. For this presentation the levels were adjusted to be approximately equal to the normal speech, and the listeners were given approximately 6 seconds for their choices.

RESULTS. Figure 7 shows a comparison graph of percentage of speakers identified correctly by the various subjects. The upper data is for the normal speech, and the lower data is for the shouted speech. The average identifications are 88.5 and 47.2 respectively.

Table I shows the confusion matrix for the normal speech. The mistakes in identity between talkers RCL, DJM, and AMN occurred because their voice quality and fundamental pitch were very similar. The first two talkers, DEB and OCJ, differed considerably in these characteristics as will be shown later in Fig. 8. Table II is the confusion matrix for the shouted speech. In this case OCJ, who has a very high normal pitch, succeeds in being recognized even while shouting. This is not the case for the remaining four talkers. Their identity is essentially destroyed.

DISCUSSION. Another interesting facet of these tests is the similarity of the fundamental pitch frequencies for the different talkers. The pitch frequencies for normal and shouted speech for the five talkers is shown in Fig. 9. The pitch frequencies were measured from speech spectrograms by measuring the frequency difference between harmonics. Although the pitch of the normal speech encompasses a wide range, the range of the pitch of the shouted speech is restricted. Also, the average pitch of the shouted speech is approximately three time higher than the pitch of the normal speech. This tripling of the fundamental voice frequency has been observed before in several of the emergency communications analyzed at Bell Telephone Laboratories.

In addition to the pitch analysis, sound spectrographs were made for each utterance of the five talkers - both normal and shouted words. Figure 9 shows the spectrograms of the talkers for the word "tan." A striking similarity in the voice pattern and speech duration exists for the normal cases. Except for the different pitch periodicities (vertical striations) the spectrograms are comparable. In the series of shouted spectrograms, however, certain obvious identity clues are observed. DEB for instance shouts with brevity. However, OCJ greatly lengthens the duration of the shouted utterance. DJM, RCL, and AMN are about equal in duration for the shouted utterance; their shouted utterance is approximately twice as long as their normal speech. RCL's voice has a coarse crackling sound while shouting and this is very evident in his speech spectrogram.

CONCLUSION

To summarize, the following main results were obtained from the tests described in this paper:

1. Extreme shouting reduces intelligibility but not as severely as expected from educated guesses.
2. Peak clipping restores the intelligibility of shouted speech nearly to that of unclipped normal speech.
3. During extreme shouting, pitch frequencies can triple.
4. Extreme shouting greatly reduces the possibilities of experienced listeners correctly identifying the shouter.
5. The addition of noise seems to be the major cause of poor intelligibility.

As far as communication systems are concerned, these results strongly imply that poor signal-to-noise ratios either before or after clipping are the prime cause of the intelligibility degradation of emergency messages. Although shouted speech is intrinsically less intelligible than normal speech, the inherent clipping encountered in most communication systems would actually improve rather than further degrade the intelligibility. However, extreme shouting greatly reduces the chances for success in identifying the person making the communication. One immediate conclusion is that people should be trained not to shout during emergencies, but this immediately assumes that the noise level at the transmitting end has not increased. During most emergencies this unfortunately is not the case so that shouting is required to exceed the higher background noise level caused by such things as fire and explosions. Perhaps

close-talking and more sensitive microphones should be used to obtain signals with improved signal-to-noise ratios. A fixed clipping threshold set to clip only shouted speech would also seem appropriate both to improve the intelligibility of shouted speech and to eliminate the distortion encountered with normal speech in most present communication systems.

REFERENCES

1. J. C. R. Licklider & Irwin Pollack, Effects of Differentiation, "Integration and Infinite Peak Clipping Upon The Intelligibility of Speech," Journal of the Acoustical Society of America, Vol. 20, No. 1, pp42-51, January 1948.

2. A. S. House, C. E. Williams, M. H. L. Hecker & K. D. Kryter, "Articulation Testing Methods: Consonantal Differentiation with A Closed Response Set," Journal of the Acoustical Society of America, Vol. 37, No. 1, pp158-166, January 1965.

RESPONSE

STIMULUS TALKER	DEB	OCJ	RCL	DJM	AMN	TOTAL ERRORS
DEB	138		1	1		2
OCJ		140				
RCL		1	121	3	15	19
DJM			10	114	16	26
AMN		1	23	4	112	28
TOTAL RESPONSES	138	142	155	122	143 →	700

NORMAL SPEECH

RESPONSE

STIMULUS TALKER	DEB	OCJ	RCL	DJM	AMN	TOTAL ERRORS
DEB	60	4	27	26	23	80
OCJ	1	128	6	1	4	12
RCL	60	6	25	28	21	115
DJM	13	8	57	34	28	106
AMN	2	20	30	14	74	66
TOTAL RESPONSES	136	166	145	103	150 →	700

SHOUTED SPEECH

1	BATH	BAD	BASS
	BACK	BAT	BAN
2	BEACH	BEAN	BEAK
	BEAT	BEAM	BEAD
3	BUT	BUG	BUFF
	BUCK	BUN	BUS
4	CAKE	CAPE	CAVE
	CASE	CAME	CANE
5	CUT	CUB	CUP
	CUFF	CUD	CUSS
6	DIG	DIN	DILL
	DIM	DIP	DID
7	DUCK	DUD	DUN
	DUB	DUNG	DUG
8	FIZZ	FIG	FIN
	FIT	FILL	FIB
9	HEAR	HEATH	HEAT
	HEAVE	HEAL	HEAP
10	KING	KIT	KID
	KILL	KIN	KICK

11	LAY	LAME	LANE
	LACE	LATE	LAKE
12	MAT	MAP	MASS
	MAN	MATH	MAD
13	PACE	PAY	PAVE
	PANE	PALE	PAGE
14	PATH	PAD	PACK
	PAN	PASS	PAT
15	PEAS	PEACE	PEAK
	PEAL	PEAT	PEACH
16	PIG	PIT	PICK
	PIN	PILL	PIP
17	PUN	PUFF	PUP
	PUCK	PUB	PUS
18	RAZE	RATE	RAKE
	RAY	RAVE	RACE
19	SALE	SANE	SAVE
	SANE	SAKE	SAFE
20	SAP	SAD	SAT
	SAG	SACK	SASS

Fig. 1: Listener response form. The subject checks which word he thinks he heard from the six.

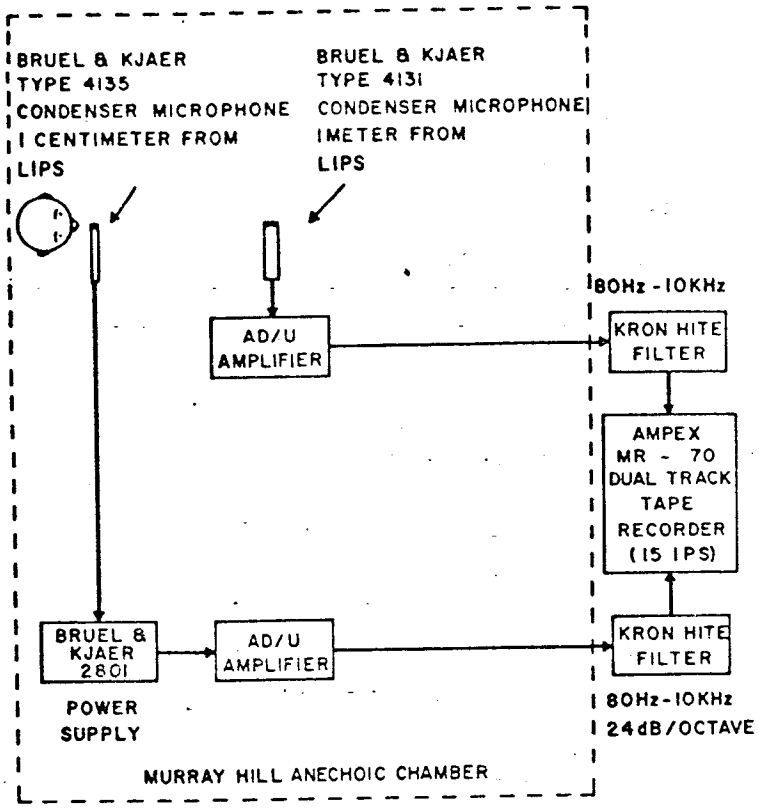

Fig. 2: Recording facility.

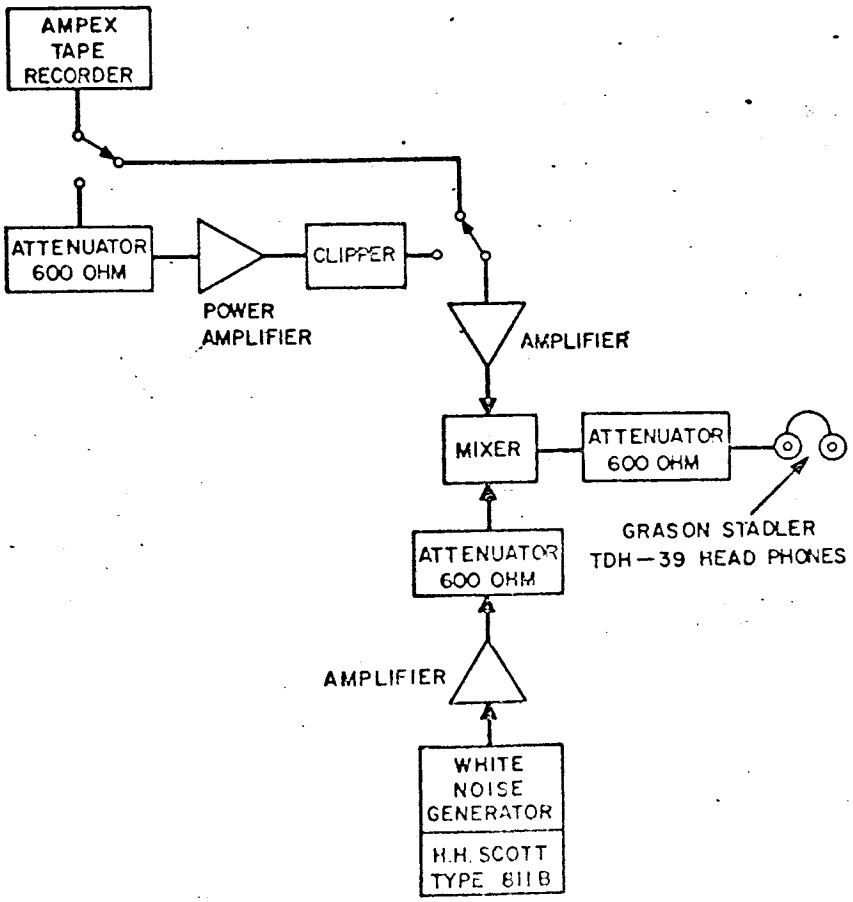

Fig. 3: Block diagram of equipment used for stimulus generation.

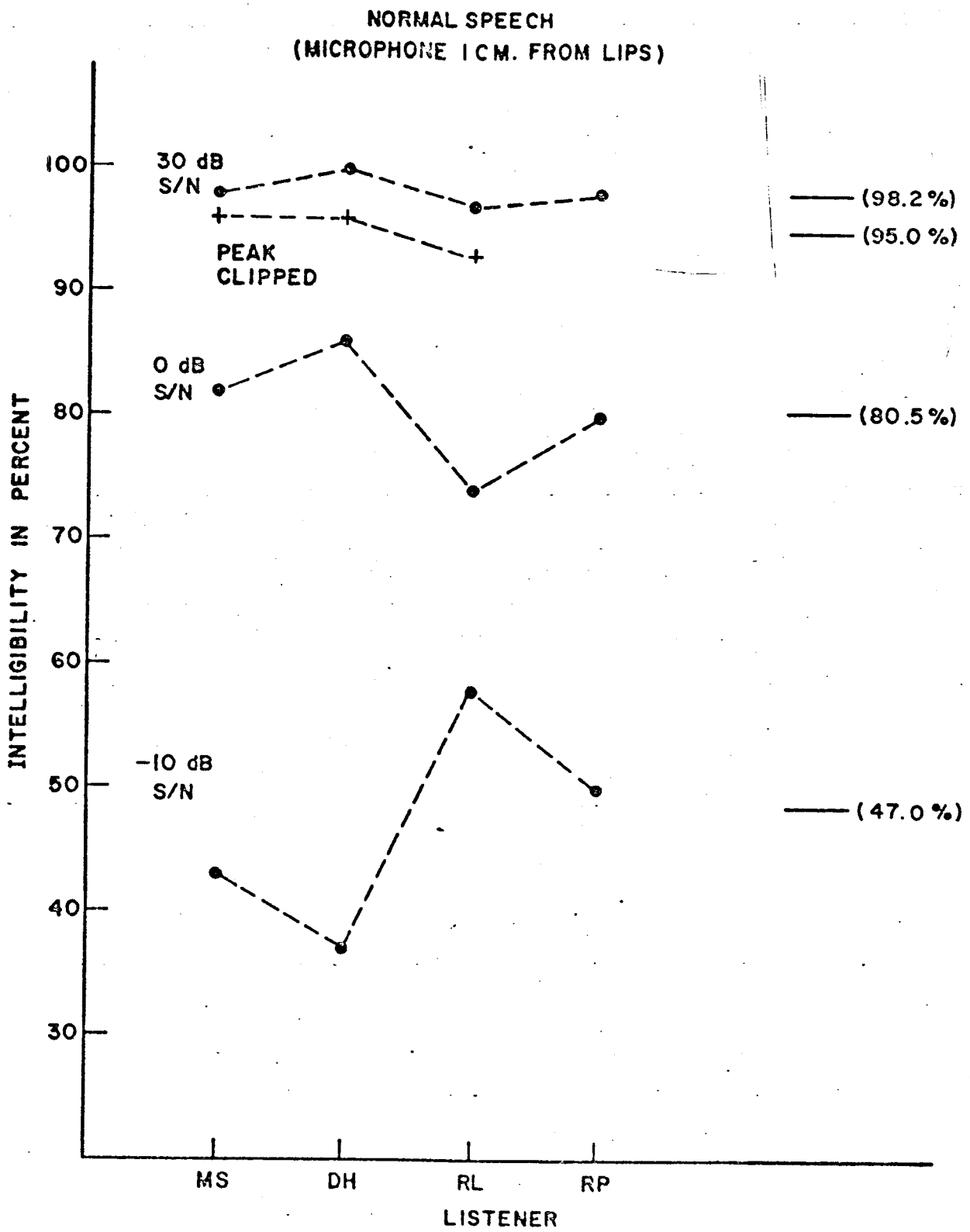

Fig. 4: Intelligibility scores (percent of words heard correctly) for normal speech.

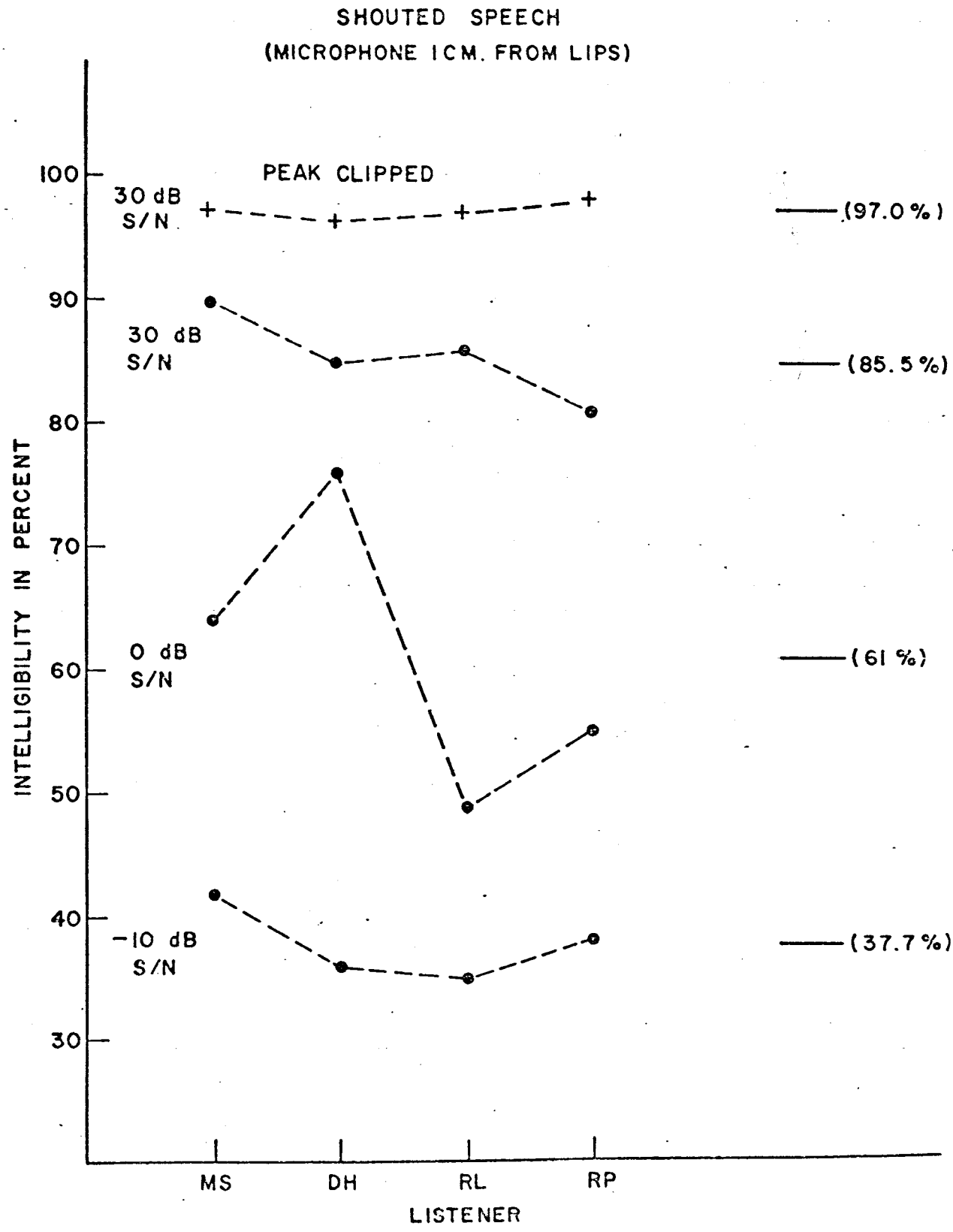

Fig. 5: Intelligibility scores for shouted speech.

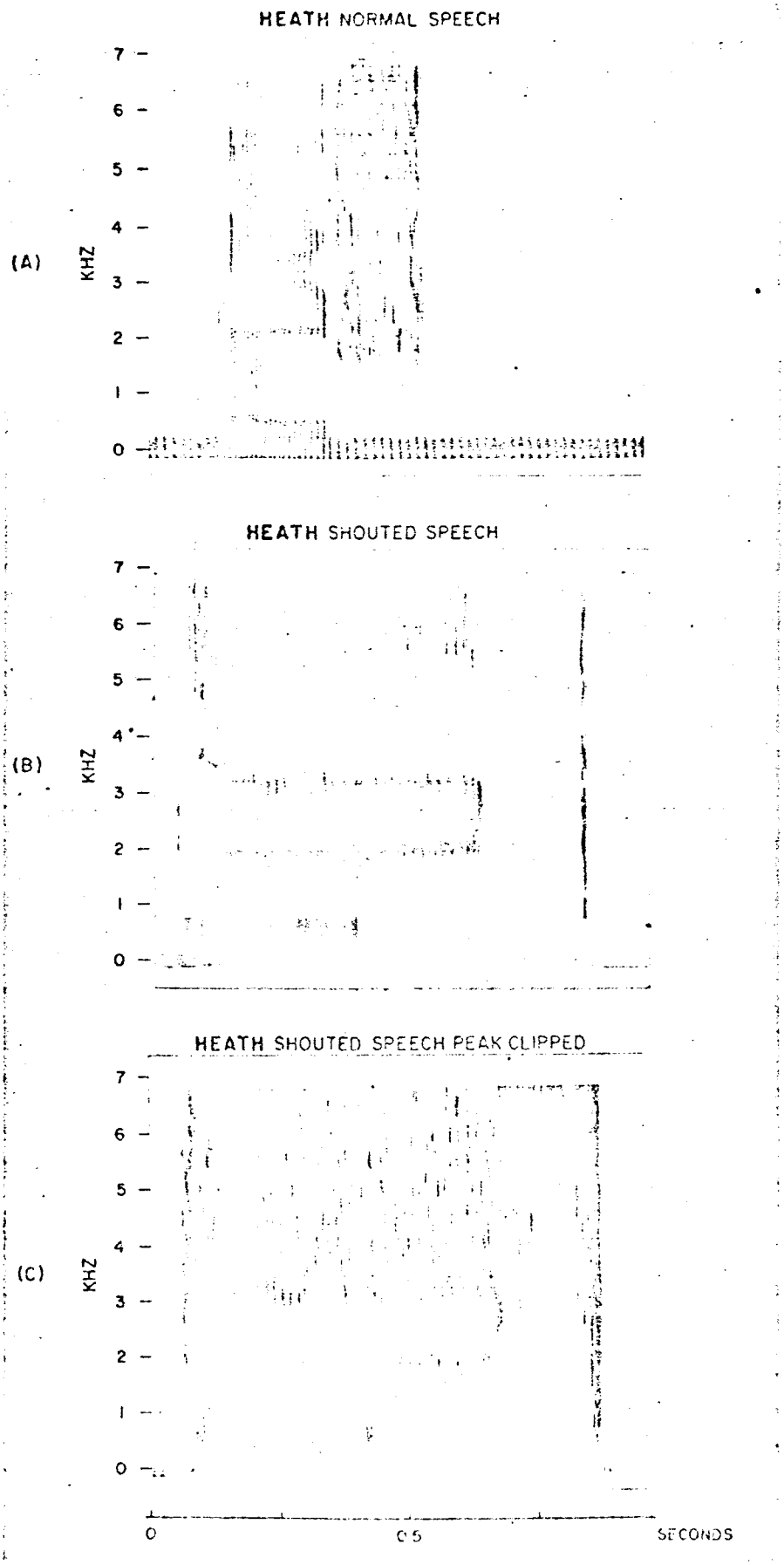

Fig. 6: Sound spectrograms of the word "heath" /i/ for normal, shouted, and clipped-shouted speech.

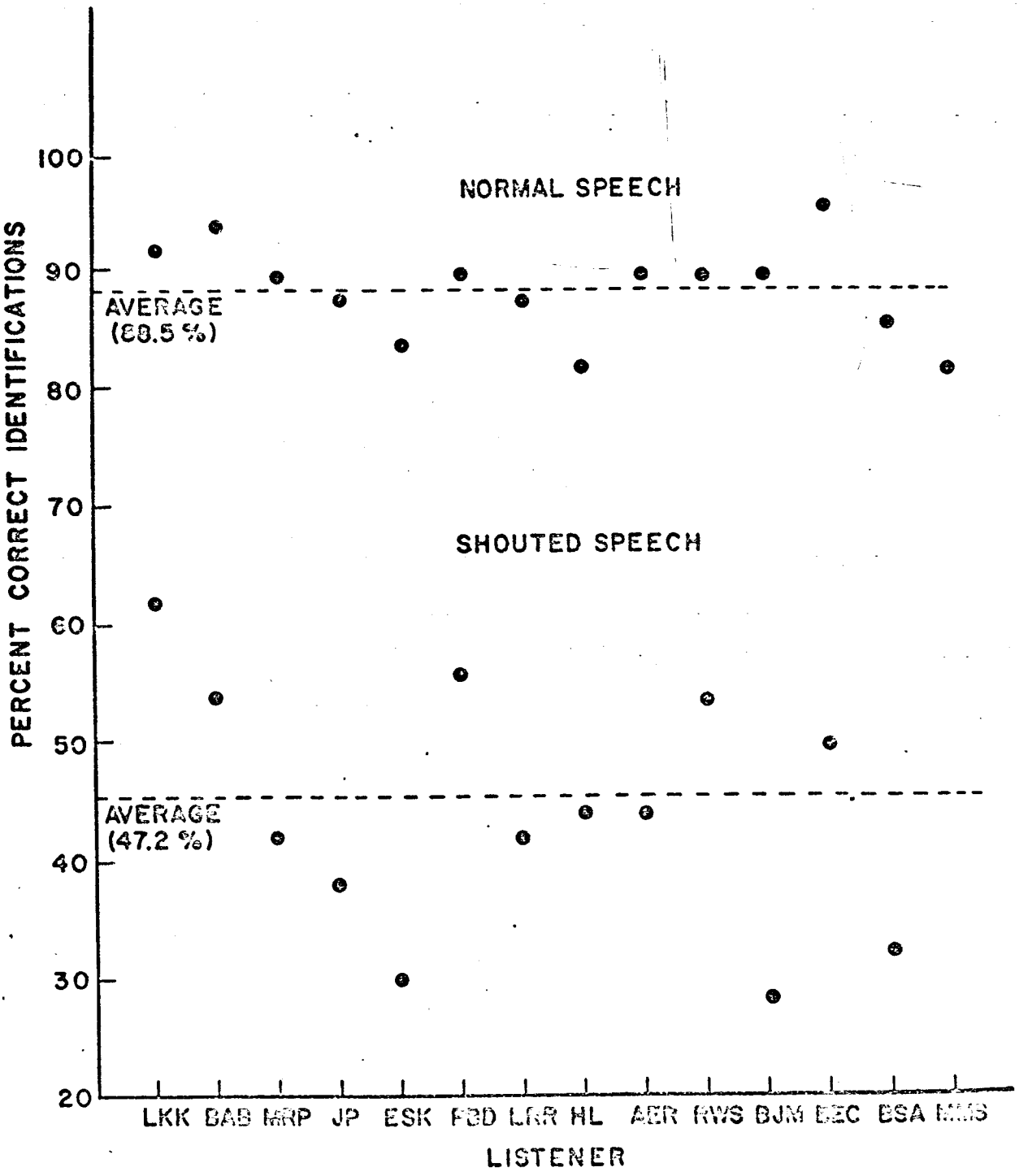

Fig. 7: Talker identification results.

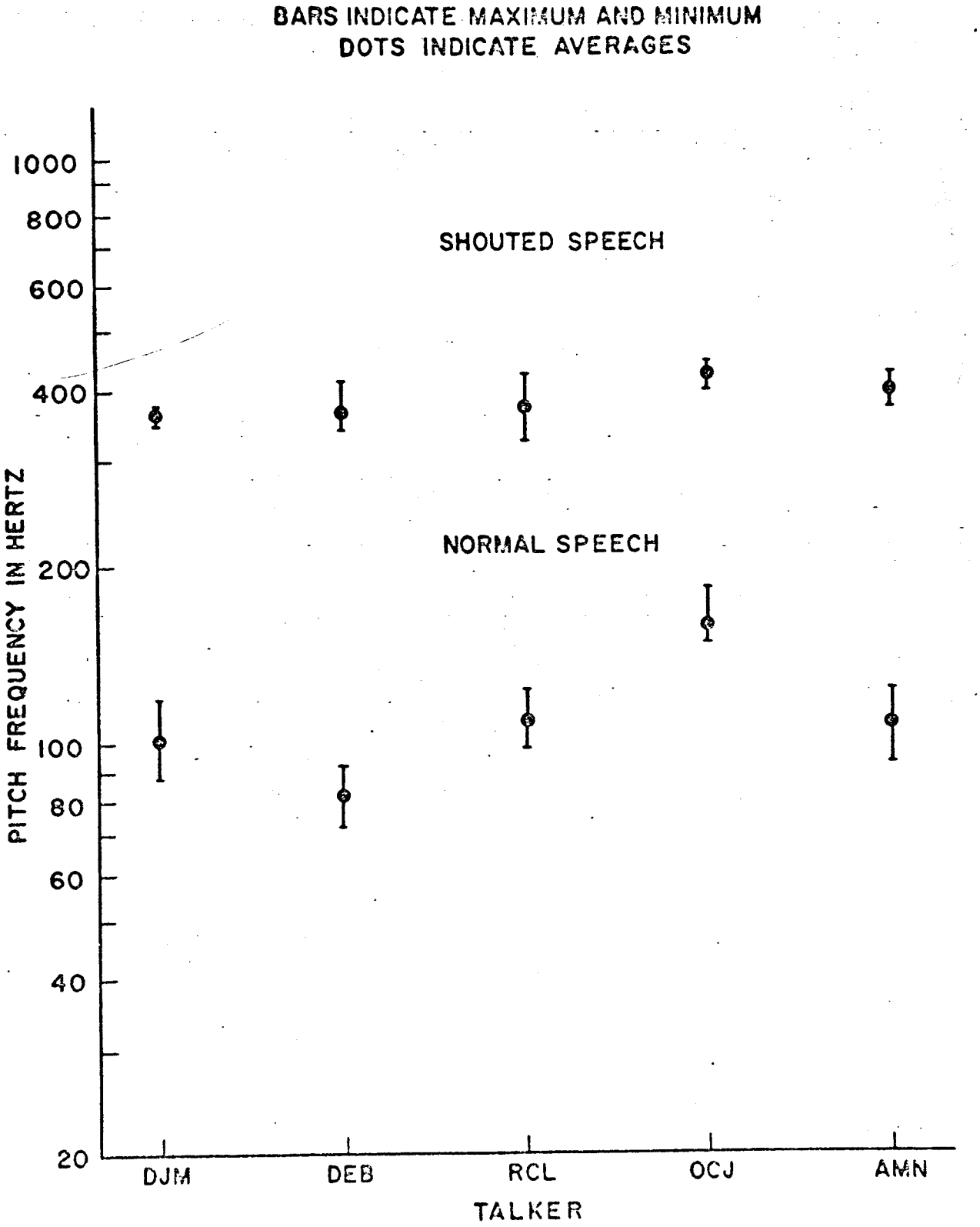

Fig. 8: Fundamental frequencies or pitch for all five talkers for normal and shouted speech.

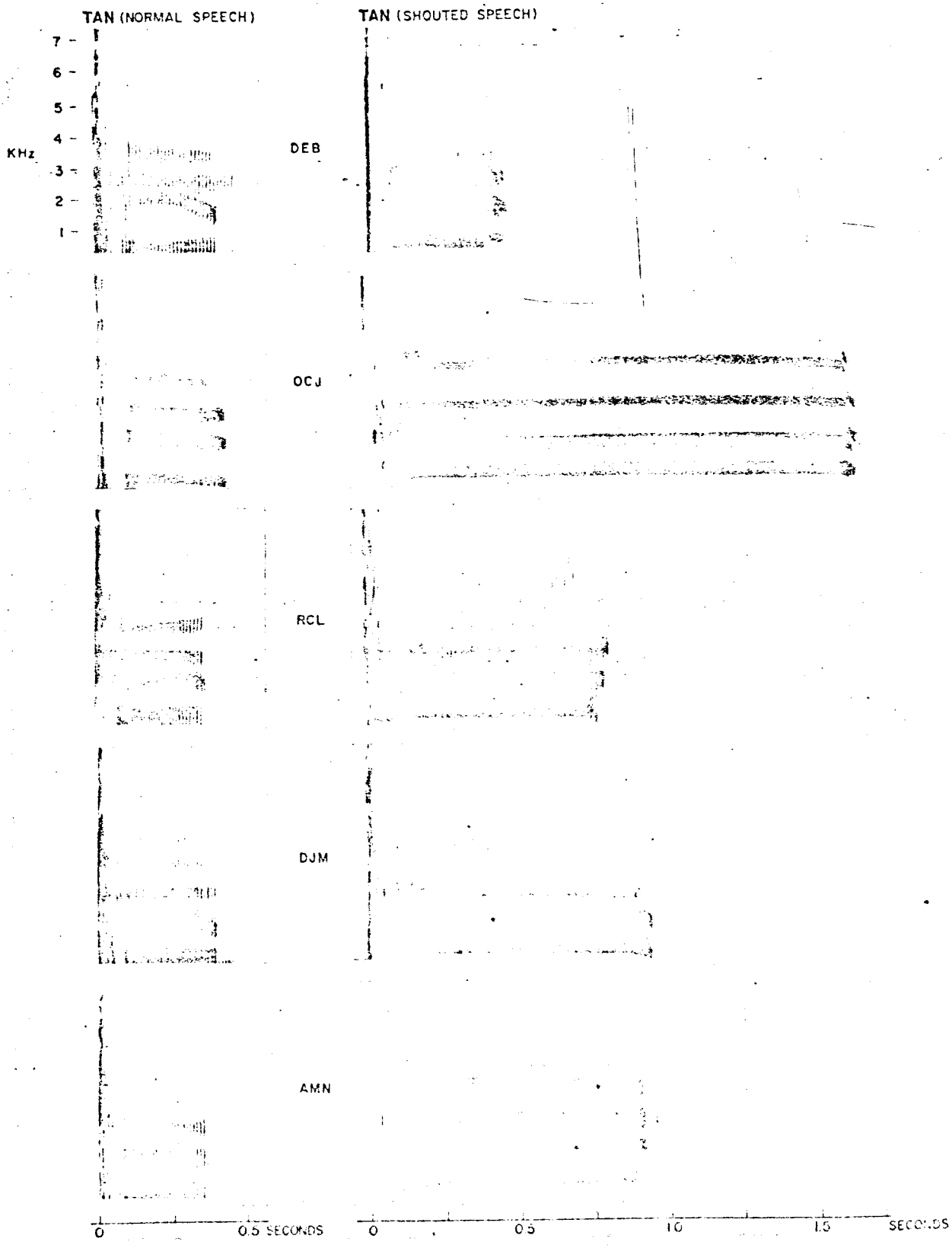

Fig. 9: Sound spectrograms of the word "tan" for all five talkers both normal and shouted speech.