Proceedings of the Institute of Acoustics

THE DECODING OF SPEECH IN RECORDINGS OF POOR SIGNAL TO NOISE RATIO

A Hirson

Speech Acoustics Lab, City University, Northampton Square, London EC1V OHB, UK.

1. ABSTRACT

Speech intelligibility may be compromised by disrupted speech production, or *post hoc* during transmission (e.g. over the telephone), recording fidelity, or by the capabilities of the listener (or measuring instrument). Factors affecting intelligibility include bandwidth, signal level, and the type and level of noise contamination. In decoding noise contaminated speech for forensic purposes, filtering may employed, before using contextual information as well as suprasegmental and segmental detail to transcribe speech material. In this paper the analysis of two recordings is presented, and the decoding process discussed in relation to contextual and phonetic detail. Digital filtering improves auditory quality, but the affect on intelligibility is not assured.

2. INTRODUCTION

Forensic tape analysis includes matching speech samples for speaker identification, the decoding of indistinct speech, and a variety of questions concerning the speaker and the tape itself [1]. Recently, methods developed in forensic phonetics have even found applications in tracing the origin of music recordings [2]. However, the high quality recordings associated with forensic musicology are relatively rare in forensic phonetics, and poor signal to noise ratios create particular difficulties in the auditory and acoustic phonetic analysis of forensic speech recordings.

This paper will review the area of speech in noise and the related subject of speech intelligibility in the light of two tape analyses. The first recording, **A**, comprised a number of tapes, all of limited bandwidth, (some of telephone bandwidth and some with a bandwidth below 1500Hz); and one distorted by variations in recording speed. The brief was to decode indistinct speech material and to compare the disputed recording with a reference speech sample for the purpose of speaker identification. Speaker identification is discussed elsewhere (e.g. see [3], [4]), and this paper will concern itself solely with the issue of transcribing speech of reduced intelligibility.

The second recording, **B**, is not strictly forensic, although similar issues arise. The recording is a digital copy of a wax graphophone recording circa 1888 of a voice believed to be that of Queen Victoria [5]. Here there is no known reference recording, and the task reduced to transcription of the barely intelligible speech.

In both recordings A and B, CEDAR (Computer Enhanced Digital Audio Restoration: [6], [7]) including the removal of clicks, hiss and distortion was used to enhance the recording. This is claimed to improve the *quality* of the recording, but quite properly CEDAR make no claims

concerning its effect on *intelligibility*. Indeed, intelligibility may be reduced. The same may be true in relation to other types of filtering, as demonstrated by the notch filtering of an intermittent steady state 400Hz tone from cockpit voice recording speech which was also heavily contaminated with broadband noise [8]. The removal of the tone was not found to improve intelligibility. Notch filtering below 300Hz, however, may improve intelligibility [9]. However intelligibility is only one aspect of auditory quality; in other words, samples with the same measured intelligibility may be adjudged to have different auditory quality [10]. This paper discusses this relationship between speech intelligibility and sound quality.

3. SPEECH INTELLIGIBILITY

Speech intelligibility has been the subject of research in several related fields and its component parts may be defined only in relation to the specific context in question. Perfect intelligibility may be taken as a one-to-one mapping between known phonetic material and the transcription by a trained phonetician or recognition device. Further refinement of the definition may involve the ease or number of passes required to minimize transcription errors.

It is clear that the measured intelligibility of the speech may be compromised at any (or all) levels between (a) speech production, (b) the acoustic environment or transmission, (c) reception or recording fidelity, or (d) transcription or other measurement instrument. In forensic speech decoding, speech production is usually relatively intact, although cases certainly exist where the intelligibility of speech production is reduced due to speech pathology, speech tempo, voice disguise, psychiatric condition, fatigue, stress, fear, or intoxication by drugs or alcohol [4].

Adverse acoustic environment, poor quality recording fidelity or transmission is also associated with decreased speech intelligibility. In all contexts intelligibility is the mapping of given phonetic material to phonetic categories known to the listener (or recognition device). In the case of disrupted production (assuming perfect transmission and ultimately a good signal to noise ratio), speech intelligibility may only be improved by gaining insights into predictable patterns of disruption and development of specialised listening skills. In cases of post-production degradation e.g. by background noise, distortion, reverberation, bandpass filtering etc., speech enhancement may improve the auditory quality of the recording without commensurate improvement in intelligibility.

An analysis of intelligibility in deaf speech attempts to define "intelligibility" in terms of 28 speech parameters which may correspond to some degree with subjective ratings of speech intelligibility [11]. The findings indicate that the strongest predictors of speech intelligibility relate to the spatial and temporal characteristics of individual phonemes (such as the contrast between the voice onset times in the pair /p/ and /b/), and to the suprasegmental encoding of contrastive word stress which assists with "chunking" of the stream of speech into word groups.

A follow-up study finds that deaf speakers may be accurately classified (83.9%) as being within one of five intelligibility quintiles (rated subjectively by normally hearing listeners) using an artificial neural net based on six segmental features [12]. This analysis is not only intrinsically interesting, it also suggests that both segmental detail and suprasegmental structure are important in resolving

disputed words. In the forensic phonetic material cited below, syllabic and suprasegmental pattern frequently provide an invaluable framework upon which narrow phonetic transcriptions may be attached, often resolving disputed words in the process.

It should be noted that subjective intelligibility ratings are the benchmark against which external measures of intelligibility are measured. There are many such measures including the SPIN (Speech in Noise) test which measures the ability of normally hearing listeners to identify test words of controlled predictability in carrier sentences [13]. This test has been used successfully to measure the performance of normally hearing listeners in identifying test words contaminated with (12-speaker) babble of varying signal to noise ratio. More sensitive assessments measure the robustness of different speech sounds under conditions of different noise levels [14].

This suggests the important differences between intelligibility used to describe pathological speech and the intelligibility of intact speech contaminated with noise. In the former, an increased in intelligibility may be achieved by improving the accuracy of production of (or sensitivity of the listener to) those features believed to comprise intelligibility. The intelligibility will be determined by the nature of speech disorder and the listening strategies adopted by the listener. On the other hand, the intelligibility of relatively intact speech will be determined by the recording bandwidth, speed variations, signal to noise ratio, and the nature of the noise. The importance of segmental and suprasegmental cues as well as contextual cues under these adverse conditions has obvious practical consequences in speech decoding.

4. CONTEXTUAL INFORMATION AND INTELLIGIBILITY

The intelligibility of speech may often be significantly improved if visual contact with the speaker is possible e.g. in analysing video recordings, since the visual information about lip and tongue movement for front speech sounds and gestures may contribute towards the overall meaning of disputed utterances. Furthermore, binaural listening strategies, locating the generally localised speech signal source in a three dimensional space from within the generally less localised noise, may be very important in free-field perception. However, such information is usually denied the forensic phonetician provided only with poor quality audio recordings.

Disputed recordings in noisy environments, often transmitted by standard telephone lines (340 - 3400Hz) and recorded on recording media with bandwidths as low as 1300Hz comprise the majority of forensic phonetic recordings. In some recordings the signal to noise ratio may be so poor as to make the differentiation between speech from noise very difficult, particularly if the noise covers the speech frequency range, is modulated in a speech-like way (rather than being steady state) or is naturalistic multi-speaker speech babble (the most confounding). In any of these adverse circumstances, a method of working downwards from larger to smaller units has been found to be an effective method of decoding.

After mapping the recording into those portions believed to contain speech, the latter is 'chunked' into words or larger units, before marking syllabic structure, stress patterns, and intonation contours.

Individual segments are then identified and described phonetically using standard phonetic notation. If words remain unresolved, a narrow phonetic transcription of disputed words often produced together with spectrographic analysis as well as extra-phonetic information provides the raw material for a final orthographic transcription. Recording A provides an example of how extra-phonetic information about a speaker's dialect may aid the transcription of a disputed word shown in Figure 1.

Line	speaker	Final orthographic transcription
1	Y	Well, W-, No, what I'm saying to you is: I want to know roughly, within a
2		couple or three or four days so that I can be around someone...
3	X	... to say you weren't there.
4	Y	That's right. [X: Right] Do you understand? I-, I-, I appreciate if you're
5		gonna get them out beforehand and put 'em away somewhere and then have
6		a moody break-in. I appreciate that.
7	X	That's- that's what's gonna happen.
8	Y	Right. Well when you gonna have the moody break in...
9	X	{SOUND OF THROAT CLEARING}...you wanna be somewhere else.

Figure 1: Context of the word *moody* from the final orthographic transcript of Recording A. (Key: interjections are marked in square brackets and non-speech sounds in curly brackets)

The intelligibility of the speech in this portion of the recording is satisfactory, but despite repeated analysis, no viable alternatives could be found for the word "moody" repeated at line 7 and 10. In resolving this word, Y's accent suggested a London Cockney dialect of English. Other parts of the transcript include words such as *Blimey* (From: *Cor blimey* = corruption of "God blind me!"; *cozzer* (amalgamation of *copper* and *rozzer* = policeman, 19C slang) and *Old Bill* (also = policeman). The repetition of the same word on two occasions suggests that the transcription (if not necessarily the spelling) is correct, and dialectal features in the remainder of the recording led to a search for "moody" in dictionaries of slang. Dictionaries of slang [15, 16] list *moody* = simulated or fake, low slang from (Cockney) rhyming slang: Moody and Sankey (late 19C American Evangelists) rhyming with "hanky panky" (trickery). A *moody break-in* is therefore a simulated burglary, entirely consistent with the context. The speakers in this recording also made extensive reference to people and places not known to the analyst, but in most cases these extra-phonetic details were readily supplied by the source agency.

Other factors affecting intelligibility of recording A were:

(a) restricted recording bandwidth
(b) distortion due to incorrect and variable speed of recording

(c) significant broadband noise and non-speech transients
(d) low signal level for distant speaker
(e) speech of one speaker masked by speech of the other.

The narrow bandwidth of the recording could not be improved, and many speech sounds particularly voiceless fricatives such as the first consonant in the words *fin, thin, shin, sin* and plosives such as the first sound in words such as *pin, tin,* and *kin,* as well as the glottal stop may be lost completely or partially. Variable (and reduced) speed was remedied manually using a REVOX reel-to-reel tape recorder and a fundamental frequency (f_0) estimation from a time-domain peak picker [17] of one speaker was used as a reference guide. Remaining distortion, as well as broadband noise and transients were attenuated by digital filtering (CEDAR: Computerised Enhanced Digital Audio Restoration).

The most straightforward method of raising the level of the distant speaker relative to the near speaker is to isolate his/her speech using a waveform editor. Subsequent increases in the level of the latter's speech does not then result in distortion of the more intense, near, speaker. Instances of both speakers talking simultaneously is the most difficult to decode, particularly when the speech characteristics of the two speakers are relatively similar.

5. THE 'QUEEN VICTORIA' RECORDING

The remarkable story of the recovery of the wax graphophone cylinder No. 1929-607 believed to contain the only extant recording of the voice of Queen Victoria is described by Tritton [5]. The recording was played for the first time in sixty years in June 1991 using a modified electronic phonograph and a elliptical stylus used for 78 records. The unfiltered recording comprises three tracks, a test track with only background noise, the first band with voice identified as male intoning a ditty containing the words: "Now how is this today", and twelve whistled notes. However, it is the second band containing barely intelligible speech that is claimed to be the voice of Queen Victoria. Despite painstaking historical work and CEDAR-filtering, this recorded speech material has not to my knowledge previously been subjected to any rigorous phonetic analysis.

The identification of the gender of the male and female speakers, as well as the direction in which the symmetrical cylinder needed to be played was based on intonation contours, and subjective evaluations of speakers' F_0 and speech tempos. Using these as guides Tritton [5] describes how The National Sound Archives played the recording in both directions at speeds of rotation of 160, 140, and 130rpm. Although the speed was not known *a priori,* it did appear to be relatively stable (<5% variation), suggesting that it was motor or mechanically driven rather than by the earlier foot treadle systems. However, my own measurements of inter-harmonic intervals (or fundamental frequency, f_0 of 260Hz-320Hz are surprisingly high, and to produce my own transcription (below) I reduced the speed yet further to give f_0 values of 200-250Hz (and a speech tempo of 4.12 syllables/second). These values are closer to normative data for (contemporary) adult females, and conversational speech tempo. The remainder of the analysis consisted of repeated listening to individual segments and groups of segments, and Figure 2 demonstrates how the final

orthographic transcription is based on the phonetic description.

Peter Copeland and others at the National Sound Archive produced a transcription, quoted in Tritton [5] as: "Greetings.......that the answer can be.....and I've never forgotten". An alternative for the beginning of the track is also quoted as: " Greeting Britons and everybody" (p93); a summary of my own transcription is reproduced in Figure 2

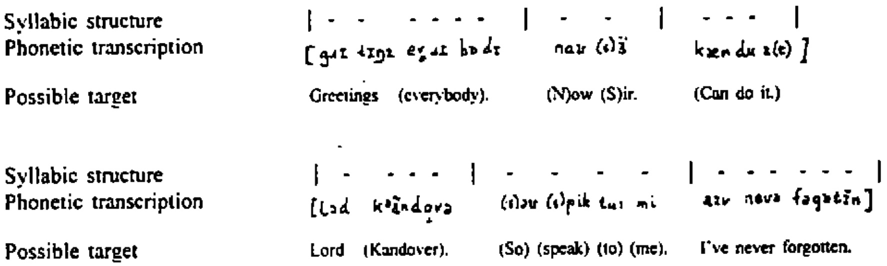

Figure 2: Syllabic, phonetic and orthographic transcriptions of Band 2. Pauses are marked with |, elements which remain in doubt even after spectrographic analysis are marked in brackets.

My transcription, summarised in Figure 2, is based on auditory and spectrographic analysis of the CEDARised version (bandwidth 587-1367Hz) together with the unfiltered version. Where discrimination between different phonetic elements was impossible, the best candidate is presented above the "best match" word in English (lower line). The portion transcribed as "everybody" is only very tentative, the speech sounds being very faint, and each syllable being strangely isolated. Similarly, the phrases "Can do it" and "So speak to me" are barely audible, and segmental detail cannot be transcribed with great certainty. In contrast, the utterance, "Lord Kandover" is transcribed with greater certainty, and it is here that further historical data might be useful.

As there is no other known recording of Queen Victoria, no reference recording is available, and speaker identification is impossible. Comparisons with speech patterns with current members of the royal family, or details of Queen Victoria's linguistic history are interesting, but ultimately of little use for speaker identification given the poor quality of the disputed recording.

6. CONCLUSIONS

In producing the final transcriptions for recordings A and B, both the CEDARised and unfiltered recordings were available. The latter were subjectively rated as having a higher auditory quality by both myself and others, but disputed words were frequently resolved by recourse to the original unfiltered version. Although filtering undoubtedly removes distracting non-speech sounds, acoustic

detail relating to individual speech sounds may inadvertently be lost and intelligibility may suffer as a consequence. In both cases presented here the unfiltered signal was found to be useful in resolving disputed segments or words.

7. ACKNOWLEDGEMENTS

I am grateful to The National Sound Archives for permission to publish work on the "Queen Victoria" tape, and to a BBC television programme for permission to reproduce the extract of recording B.

8. REFERENCES

[1] J BALDWIN & J P FRENCH, 'Forensic Phonetics', London: Pinter Publishers (1990)

[2] D M HOWARD, A HIRSON, & G LINDSEY, 'Acoustic Techniques to trace the origins of a musical recording', *Journal of the Forensic Science Society*, 33 (1) p33-37, (1993)

[3] F NOLAN, 'The Phonetic bases of speaker recognition', Cambridge: Cambridge University Press (1983)

[4] H HOLLIEN, 'The Acoustics of Crime, The New Science of Forensic Phonetics', New York: Plenum Press (1991)

[5] P TRITTON 'The Lost Voice of Queen Victoria', London Academy Books Ltd.(1990)

[6] S V VASEGHI, P J W RAYNER & L STICKLES, 'Digital signal processing methods for the removal of scratches and surface noise from gramophone record', *Image Technology (Journal of the British Kinematograph Sound & Television Society)* 69 (10) p457 - 461 (1987)

[7] S V VASEGHI & P J W RAYNER, 'The effects of non-stationary signal characteristics on the performance of adaptive audio restoration systems', *IEEE-ICASP*, 1 p377-380 (1989)

[8] A HIRSON & D M HOWARD, 'Issues arising from the spectrographic analysis of a Cockpit Voice Recorder tape', *Journal of Language and the Law*, 1 (1994) (In press).

[9] R H WILSON, J P Preece & C S CROWTHER, 'Enhancement of word-recognition performance with a filter4ing technique', *Journal of Speech and Hearing Research*, 34 p1436-1438 (1991).

[10] L H NAKATANI & K D DUKES, 'A sensitive test of speech communication quality', *Journal Acoustical Society America.*, 53, (4) p1083-1092 (1973)

[11] D E METZ, N SCHIAVETTI, V J SAMAR & R W SITLER, 'Acoustic dimensions of hearing impaired speakers' intelligibility: segmental and suprasegmental characteristics', *Journal of Speech and Hearing Research*, 33 p476-487, (1990)

[12] D E METZ, N SCHIAVETTI & S D KNIGHT,' The use of artificial neural networks to estimate speech intelligibility from acoustic variables: A preliminary analysis', *Journal of Communication Disorders* 25 p43-53, (1992)

[13] D N KALIKOW, K N STEVENS, & L L ELLIOT, 'Development of a test of speech intelligibility using sentence materials with controlled word predictability', *Journal of the Acoustical Society of America*, 61 (5) p1337-1351, (1977)

[14] R K KOUL 7 G D ALLEN, 'Segmental Intelligibility and Speech Interference thresholds of high-quality synthetic speech in presence of noise', *Journal of Speech and Hearing Research*, 36 p790-798 (1993).

[15] P BEALE, 'A concise dictionary of slang and unconventional English', London: Routledge (1992)

[16] J MORTON, 'Lowspeak: A dictionary of criminal and sexual slang', London: Angus and Robertson, Publishers (1989)

[17] D M HOWARD & A J FOURCIN, 'Instantaneous Voice Period Measurement for Cochlear Stimulation', *Electronic Letters*, 19 (19) p776-778 (1983)