The acoustic analysis of early speech recordings

Michael Ashby
Speech Hearing & Phonetic Sciences, University College London (UCL), UK
m.ashby@ucl.ac.uk

Abstract: The shortcomings of early phonographic recording have never been systematically evaluated in relation to the specific acoustic properties of speech. This paper reports preliminary investigations into limitations on the quantitative analysis of recordings which were made prior to the introduction of electrical recording in 1925–1926. Analysis of such early recordings would be valuable in historical sociophonetics, for example, or in the replication of early speech science research which utilized recordings. To provide some foundation for quantitative study, we therefore attempt to determine the characteristics of acoustic recording systems from general considerations, from contemporary analysis and evidence, and from modern spectral analysis of historic recordings. It emerges that major deficiencies in the frequency response achieved with acoustic recording severely hamper spectral analysis (e.g., for vowel formant frequency estimation). Proposed future work will attempt direct measurements on surviving or replicated components of acoustic recording systems.

1 Introduction
The Edison phonograph was announced in 1877, and for the next 50 years all recordings, whether on cylinder or disc, were made with an entirely acoustic (mechanical) process (Figure 1). Sound entered the apparatus through a recording horn, \( h \), which for a single speaker or performer might be around 1 metre in length and 25 or 30 cm in diameter at the mouth. The narrow throat of the horn led to a sound-box containing a circular diaphragm \( d \) around 30–50 mm in diameter, made from glass or mica, which was supported around its circumference but free to vibrate at the centre. A stylus, \( s \), commonly of sapphire, its point bearing on a steadily rotating cylinder or disc of suitable wax, was linked to the other side of the diaphragm in such a way that the vibrations either cut a groove of varying depth (the ‘hill-and-dale’ process) or imposed lateral excursions on a groove of constant depth (as in Figure 1). The cutter was made to advance at a constant rate across the surface, producing finely spaced grooves on the cylinder or disc. Valuable accounts of various aspects of the acoustic recording process are given in numerous papers by Brock-Nannestad (e.g., [3]–[6]).

Around 1914, researchers at Bell Laboratories began ‘a comprehensive survey of speech and hearing to obtain the fundamental facts on which to base the design of apparatus and systems for telephone use’ [8, page v]. Among the outcomes of this very productive project was a systematic analysis of all the components of a phonographic recording system, and a series of
patents specifying an electrical system that represented a step change in the fidelity of recorded sound, described in a celebrated paper by Maxfield and Harrison [16]. The new system was licensed for use to record manufacturers in 1925, and the first commercial electrical recordings were released towards the end of that year. The superiority of electrical recordings was evident to all who heard them, even when (as often) they were played on the acoustic reproducers then in common use. Within a very short time, all commercial record manufacturers adopted electrical recording. Though the earliest have as yet been little utilized by researchers, commercially produced speech recordings from 1926 onwards can yield satisfactory acoustic analyses with standard digital tools [21]. Outside the commercial context—for example among field-workers or in dialect archives—acoustic recording continued for a few years, but by the mid-1930s a number of manufacturers were marketing fairly portable direct-to-disc electrical recording systems which supplanted it [9], [10].

2 Early speech recordings

The consequence of the history just outlined is that recordings from the first fifty years of recording technology were made with systems of which the detailed characteristics are largely unquantified. At the same time, these earliest recordings are of great historical interest. They potentially provide the earliest audio sources for historical linguistics, dialectology and sociophonetics, while certain recordings are themselves directly relevant in the history of phonetic science.

Early sound recordings interface with the history of speech communication research in numerous ways [23]. On the application of early sound recording techniques as an integral part of the instrumental methodology in phonetic research, see [14] and [5]. In many such cases the original recordings no longer survive or cannot now be replayed, particularly if made with non-standard apparatus. Once commercial standards for cylinder and disc recording emerged, the chances of recordings surviving in a playable form became much greater.

Of course, the utility of sound recordings in the history of phonetic science does not always depend critically upon quantitative analysis of the signals they contain. An obvious example is oral history as recorded by teachers and researchers (e.g., [2]), where the content, rather than the signal itself, is the primary focus of attention. Similarly, early recordings made to survey languages or dialects have great value as linguistic documents even if they have not been submitted to (or will not permit) detailed acoustic analysis. But it is true that almost any recording can yield more information if acoustic analysis is possible. With the passing of time, what begins as, say, a contemporary language-teaching course may become more valuable instead as evidence for the standard pronunciation of the past [21].

While considerable effort has been devoted to the question of restoring and digitising early recordings of musical performances, the same cannot be said of speech. For example, despite the promising title *(Listening to the past)* of a recent edited collection [11]), none of the contributors to that volume attempts acoustic analysis with any speech material that predates electrical recording, and the technical problems posed by old recordings are not addressed.

It therefore seems desirable to investigate the characteristics of acoustic recordings in more detail, to determine what analysis procedures might appropriately be applied to speech recordings of historic interest.

3 General considerations

3.1 Horn and diaphragm

The two main components of an acoustic recording system are the horn and the diaphragm. The use of horns both to gather and disperse sound has been known since antiquity and horns form part of many musical instruments, though the design of horns was largely empirical before
mathematical analyses appeared in the twentieth century [1]. The horn is necessary in acoustic recording as an impedance transformer, permitting feeble airborne vibrations to move the relatively heavy diaphragm. This task it performs well, but the horn also has undesirable characteristics. First, it exhibits a series of resonances—the fundamental resonance, for a recording horn of typical dimensions, being of the order of a few hundred Hz. Second, the horn acts increasingly as a high pass filter as the wavelength of incident sound grows in relation to the size of the mouth of the horn. A smaller horn (for example as used with a portable cylinder recorder in the field) can be expected to exhibit greater low-frequency loss than the large horns found in the recording studio [7, page 257].

The diaphragm, essential as the component which converts airborne pressure variations into mechanical movement at the cutter, likewise exhibits its own resonances. The relatively low principal resonant frequency of any practical diaphragm, and the mass of the cutter mechanism attached to it, account for the limited high frequency response achieved.

Neither the horn nor the diaphragm resonances were directly damped in the design and construction of acoustic recorders, though the diaphragm must have been damped to some extent by the movement of the cutter through the recording wax. In any case, with no amplification available, there was little scope for damping since it would have lowered both recording and reproduction levels unacceptably. The tapes resembling bandages which are seen fixed around the recording horn in many early photographs served to suppress entirely separate bell-like resonances caused by transverse waves in the material of the horn (which was typically thin sheet metal).

It is worth noting too that reverberation in the recording environment was not damped. On the contrary, it was to a degree actively promoted as a valuable source of sound reinforcement. Some early speech recordings appear indeed to have a perceptible echo, sounding as though they have been made in an unnecessarily large empty studio. Plainly, if reverberation is present, it might be a complicating factor in attempts to measure time intervals such as VOT, or transients such as plosive bursts.

3.2 Noise

The noise we are accustomed to hearing in the reproduction of old recordings has multiple origins. So-called ‘cutter noise’ was generated as the recording stylus encountered granularities in the wax blank and as swarf separated from the cutter. This noise was audible to performers as they made the recording, the recording horn acting equally well as a reproducer. The compound from which the eventual pressings were made—so-called ‘shellac’ in the case of most commercially-produced discs—also has a granular surface, and further noise is therefore contributed when the reproduction stylus traverses these irregularities. Recordings are subject to deterioration through wear (especially if replayed with inappropriate equipment) and, it is believed, through damage to the surface resulting from environmental factors during storage such as seasonal annual temperature cycling, and hydrolysis. It is worth pointing out that unlike the wanted speech signals themselves, the unwanted surface noise (or at least that component of it generated during the reproduction process) has a very wide (essentially unlimited) bandwidth. This is an argument for making digital transfers of old recordings at much higher sampling rates than would appear at first glance to be required, since algorithms designed to reduce noise, crackles and clicks are more likely to work effectively if the unwanted sounds have been adequately sampled.

3.3 Rotation rates

Motive power for recorders and reproducers came at first from direct manual cranking (with or without some form of governor), and then from spring motors, electric motors, or gravity mo-
tors driven by descending weights. For commercial cylinder and disc recordings, various standard rotation rates were nominally in use, though the precision with which any intended standard was achieved varied considerably [19]. Fieldworkers using cylinder recorders might deliberately reduce recording speed so as to fit a performance into the capacity of a single cylinder. Musicologists understood the value of recording a reference pitch (from a pitch pipe) at the beginning of a field recording, but the practice was not followed with speech recordings.

In practice, uncertainty over reproduction rate is not normally a major stumbling-block to analysis, even when fundamental frequency (pitch) is the primary target of the investigation. Whether the concern is with intonation, or with lexical tone, it is generally not absolute pitches but rather pitch intervals (ratios) that are of interest. Speed differences result in a global scaling of frequency, but leave pitch intervals and pitch patterns unchanged when viewed on a logarithmic scale.

In fact, both the high level of noise and the deficient low-frequency response of older recordings are bigger problems than speed in quantifying the fundamental, since they hamper the operation of pitch-tracking algorithms on which the automatic plotting of fundamental frequency patterns depends. High-amplitude noise impulses are easily mistaken for glottal excitation pulses, and pitch trackers always incorporate, or work in conjunction with, a voicing detector, which may be defeated by signals in which the lowest harmonics of the voice are greatly attenuated, or missing.

4 Frequency response characteristics

The most significant unknown which impedes the quantitative acoustic analysis of early speech recordings is the overall transfer function of the acoustic recording system. Various approaches can be used to throw light on this.

4.1 Contemporary evidence and testimony

The main shortcomings of the phonograph—lack of both high and low frequencies—were known from the outset in general terms. Alexander Graham Bell had lodged his first patent for the telephone in 1876, just one year ahead of Edison’s announcement of the tinfoil phonograph, and it was natural that the two talking devices would be compared. As early as 1878 Preece and Stroh show an excellent grasp of the general acoustic properties of various categories of speech sound, and an understanding that the limited speech intelligibility offered by both the early telephone and the phonograph resulted from similar deficits in the two devices, particularly the failure to reproduce high frequency energy [20].

It was appreciated too that the low-frequency response of the phonograph was deficient. Rayleigh noted a discrepancy in the reported amplitude of the fundamental between studies done with live and recorded vowels (22, page 477). Listening with the aid of an acoustic resonator, he confirmed that the first harmonic is strong in live spoken examples of vowels, though apparently largely absent in the phonograph recordings of his day.

The characteristics of diaphragms and horns were investigated empirically in great detail by D. C. Miller at least as early as 1912 [17]. The test signals were supplied by a bank of 80 specially commissioned organ pipes tuned in semitones to cover the range 129 Hz–12400 Hz and judged to be of equal loudness throughout. Miller showed that both the horn and the diaphragm exhibit a series of resonances. The frequencies of the horn resonances are chiefly dependent upon its length, though the diameter and the flaring of the mouth also have important effects. The principal resonance of the diaphragm is greatly affected by the housing in which it is enclosed and by the type of mounting around its edge, while some of its higher resonances may be inharmonic, indicating that portions of the diaphragm are vibrating independently. The
response of a particular horn and diaphragm system is shown in Figure 2.

Figure 2. Frequency response of a horn and diaphragm combination, as determined by Miller [17]. Curve b is the composite response, a (dotted) the response of the diaphragm alone; c is the ideal response for equal loudness. Note that the amplitude scale a is linear, not logarithmic.

Miller’s investigations were directed at the improvement of his precision mechanical-optical oscillograph, which he termed the ‘phonodeik’. But as he notes, ‘[t]hese same difficulties are present in the reproductions of the talking machine and the telephone, perhaps in a doubled degree since they may occur once in making the record and again in reproducing it’. There is no reason in principle why Miller’s methods should not have been immediately extended to an investigation of the phonograph, potentially leading to considerable improvements in acoustic recording and reproduction, but apparently this did not happen. In fact, no comprehensive contemporary attempt to determine the frequency response characteristic of an acoustic recording system has hitherto come to light. Surprisingly, Maxfield and Harrison, who might be thought the most likely researchers to have done this, give the characteristics of two acoustic reproducers but not of a recording system.

It is not clear whether the manner in which the recording horn contributes to the low-frequency loss was fully understood at this period. Miller does note that ‘the response below the fundamental of the horn is very feeble’ [18, page 159]. His remedy is to use a long horn: ‘For the study of vowel sounds, the horn employed has a length of about 48 inches, giving a fundamental frequency of about 125’. This is probably rather larger than the horns commonly used at that time for speech recordings in the studio, and certainly much larger than the horn of a cylinder recorder suitable for use in the field.

4.2 Evidence from noise spectra

Brock-Nannestad [4, page 41] noted the ‘formant-like’ coloration of the background noise (distinct from surface noise) heard from unmodulated portions of a recording (such as the lead-in grooves), and reasoned that the spectrum of this must be closely related to the transfer function of the recording apparatus (because the vibration of the stylus in response to the mechanical shocks it receives during the cutting process will have been constrained by the resonances of the system of which it is a part, including the horn). The example he gives certainly shows a very irregular response. A British patent [3] gives further detail and background, and develops proposals for inverse filtering.

4.3 Evidence from long-term spectra of recordings

Another approach suggested itself to the author of this paper—based on long-term spectra of recordings with known characteristics. For example, if the long-term spectrum obtained from an acoustic recording of a male speaking English prose were compared with a spectrum obtained from a modern recording of a similar speaker and similar content (made with a system having an essentially flat response), then subtracting one from the other should yield a fair
estimate of the transfer function of the acoustic system. Some preliminary results obtained with this method are given in Figure 3.

Unsurprisingly, it emerges that the idea is not novel, having been explored by Stockham et al. in the early days of digital signal processing [24], though the application they made was to recordings of musical performances. Stockham et al. referred to the task being attempted as ‘blind deconvolution’ since both the wanted signal and the transfer function with which it is convolved are unknowns. This term has passed into general use in signal processing.

4.4 Evidence from replication studies
A further possible source of evidence is to make measurements on surviving or replicated components of an acoustic recording system [6]. This is the most direct method, but also the most expensive and difficult to arrange. In the ideal case, a complete acoustic recording system would be operated in an anechoic chamber and evaluated by means of sweep-tone measurements. Even where complete replication is impracticable, much might be learned from similar tests conducted with components of the system—for example, on the recording horn alone. Replication studies are further considered below in section 6, ‘Directions for future research’.

5 Vowel formant frequency tracking
Several freely-available analysis packages now offer automatic formant frequency extraction, facilitating the plotting of vowel qualities in a two-dimensional formant space, which in consequence has become a routinely-employed tool in many applications of phonetics, including language description, and sociophonetics. It would be valuable to be able to extend the same technique to the oldest recordings. The attempt cited here refers to Daniel Jones’s original (1917) recordings of the Cardinal Vowels [13].

Jones’s recordings of the Cardinal Vowels constitute the most notable attempt ever made to render the classification and symbolisation of vowel qualities repeatable across observers.
Jones recorded his framework of reference vowels on three occasions—in 1917, 1943 and 1956—and it would be of interest to make a quantitative comparison of vowels from the three dates. Naturally, the 1917 recording, made for The Gramophone Company (HMV) was an acoustic recording. Admittedly, it is an excellent example of good acoustic recording, but of course it is necessarily characterised by the shortcomings outlined above.

A digitised copy of the 1917 recording was obtained from the British Library, and analysed with the *formanal* program, a formant tracker based on LPC polynomial roots and dynamic programming [12]. For comparison, the same procedure was followed with corresponding sections of the 1956 recordings.

Mean estimates of F1 and F2 for all eight of the 1917 Cardinal Vowels, and comparable mean values from the 1956 recording are plotted together in Figure 4. It is evident that the two sets of estimates are widely divergent, only agreeing reasonably well for Cardinal Vowel 4. In every other case, F1 of the 1917 vowels is considerably too high. The agreement in F2 is better, but for the high vowels [i] and [u] an unexpected formant has been identified around 1600–1800 Hz, causing those vowels to be allocated bizarre locations in the F1/F2 space.

We may ask to what specific shortcomings of the acoustic recording system the severe distortion of the 1917 formant frequency pattern may be attributed, and whether there is any prospect of compensating for them, for example by means of an inverse filter, as proposed in [3], [24].

![Figure 4](image)

*Figure 4.* F1/F2 plot comparing the Cardinal Vowels from the 1917 and 1956 recordings. The connecting lines draw attention to the approximately triangular vowel space formed by the estimates from the 1956 recording. The vowels, in order 1–8, are [i e ɛ a ɑ ɔ o u].

Examination of narrowband spectrograms from the 1917 recordings indicates that while they have a respectable frequency range for their time, there is essentially no energy below 250 Hz, with the consequence that the second harmonic of Jones’s voice is the lowest component present. Generally, too, the second harmonic itself is weaker than the third.

Radical attenuation of the lowest frequency components of the voice can be expected to have a major impact on the frequencies of the poles required to model the spectrum—that is, the ‘formants’ will no longer appear at the expected frequencies. The effect is likely to be worst for F1, and will particularly affect high vowels where F1 is low in frequency and thus nearest
to those harmonics with disturbed amplitudes. This is exactly what is found in the present investigation. This defect, as explained above, results from a steep low-frequency cut introduced by the horn. Although the shape of this cutoff might indeed be determined, there appears little hope of compensating for it so as to restore the vowel spectrum, since the signal-to-noise ratio at the lowest frequencies is so unfavourable (essentially, there is simply no low-frequency energy present to be boosted).

On the other hand, it seems possible that the unexpected formant noted in close vowels [i] and [u] might result from a spurious resonance within the recording system. There is a much better prospect of dealing with this, by designing an inverse filter to regularise the response within the limited passband of the system (say 300–3000 Hz), although success in this depends critically on determining the transfer function as set up for the particular recording under consideration.

In short, formant analysis of early recordings seems to be hampered by a combination of systemic and particular shortcomings. The particular shortcomings (evident as peaks and troughs within the passband) can perhaps be overcome, at least in part, if one or other method is successful in determining the transfer function of the system used in recording. But the systemic shortcomings (effective absence of both high and low frequencies) cannot be remedied in any simple way.

It is noteworthy, however, that human listeners can apparently make subtle phonetic judgements about vowels from recordings where the formant pattern is badly distorted. After all, for a generation or more the Cardinal Vowel system was successfully taught and learned from the 1917 recordings—and furthermore those recordings will generally have been played on small acoustic gramophones which introduced numerous additional defects of their own into the spectrum.

6 Directions for future research

Various researchers have previously described studies or projects which aim to re-create the acoustic recording process in whole or in part [6, 15]. But it seems that hitherto all researchers have had the goal of restoring historic recordings of musical performances. The specific requirements of speech recording have never been explored.

There is a case, therefore, for conducting a replication study utilising speech. A number of considerations suggest that a re-enactment of the 1917 Cardinal Vowel recordings would be a rewarding choice. The date and circumstances of the original recording session are known, and this, in conjunction with surviving archive material held by the original recording company (EMI) can guide appropriate choice of horn and soundbox (original or replicated). There is a pool of trained speakers, both male and female, who can reproduce the required vowel qualities with great precision, and potentially also provide informed insight into the vocal accommodations which the acoustic recording situation requires. The same speakers would be comfortable with the use of simultaneous electro-glottography, nasal accelerometry and ultrasound to monitor their performance in some recording sessions.

The study would be carried out in an anechoic chamber, and exhaustive sweep-tone and other measurements conducted at every stage on the recording set-up and its individual components. While it would certainly be desirable to mimic the whole recording process, including the cutting of a wax record, that final expensive stage is not strictly necessary to gain most of the advantages of the study. For the most part it will be sufficient to use microphones and miniature accelerometers placed in and around the acoustic system. A signal derived from the motion of the cutting stylus will provide a (noise-free) analog of the resulting acoustic recording. The results could also be used to test the various ‘blind deconvolution’ proposals more directly than ever before, and—if recording to wax were indeed attempted—the effectiveness of Brock-Nannestad’s proposal to utilize the spectrum of noise from unmodulated grooves.
Since the content of the recording is in its nature a systematic sampling of the whole range of vowel qualities, the results would be of wide application and could help calibrate quantitative measurements on vowels across countless other recordings from the era.

7 Acknowledgments
I am grateful for generous guidance and advice from George Brock-Nannestad in the preparation of this paper. Any errors are the author’s responsibility.

8 References


