Perception of music performance on historical and modern commercial recordings

Renee Timmers[a)]
Centre for the History and Analysis of Recorded Music, Department of Music, King's College London, Strand, WC2R 2LS, London, United Kingdom

(Received 15 July 2006; revised 10 July 2007; accepted 22 August 2007)

Performing styles as well as recording styles have changed considerably within the 20th century. To what extent do the age of a recording, the unfamiliarity with performing style, and the quality of a reproduction of a recording systematically influence how we perceive performances on record? Four exploratory experiments were run to formulate an answer to this question. Each experiment examined a different aspect of the perception of performance, including judgments of quality, perceived emotion, and dynamics. Fragments from *Die junge Nonne* sung by famous singers from the start, middle, and second half of the 20th century were presented in a noisy and clean version to musically trained participants. The results show independence of perception of emotional activity from recording date, strong dependence of perceived quality and emotional impact on recording date, and only limited effects of reproduction quality. Standards have clearly changed, which influence judgments of quality and age. Additionally, changes restrict the communication between early recorded performers and modern listeners to some extent as shown by systematically smaller variations in communicated dynamics and emotional valence for older recordings.
 [DOI: 10.1121/1.2783987]

PACS number(s): 43.75.St, 43.75.Cd, 43.38.Md [DD] Pages: 2872–2880

I. INTRODUCTION

Performing style has changed considerably within the 20th century as comparisons between performances on modern and historical recordings suggest. For example, the use of extreme tempo fluctuations as observed in early 20th century recorded performances is now seen as highly inappropriate. Additionally, the use of pitch glides has become much less common among singers and violinists, while, on the other hand, vibrato has become more prominent as an expressive device (e.g., Day, 2000; Philip, 1992).

Similarly, the conditions of recordings and the quality of recording and reproduction of sounds have changed dramatically. Registration material changed for example from tin foil, to wax, to magnetic tape. Recording horns were used in different sizes and shapes. These were later replaced by microphones and electrical amplification before the introduction of the stereo microphone. Reproduction material and equipment also changed dramatically from the use of cylinders, vinyl, and shellac discs to microgroove discs, tape, and CDs, as did the equipment used to replay them. Moreover, technical improvements influenced the recording and reproduction of sound at every stage of these developments. This resulted in considerable changes in, e.g., the recorded frequency range, the noisiness of recordings and reproductions, the recorded acoustics, as well as the balance between different voices (see, e.g., Gelatt, 1956; Copeland, 1991; Day, 2000).

What does this imply for our (current-day) perception and evaluation of performances on record? To what extent do the age of a recording, unfamiliarity with performing styles, and the quality of a reproduction of a recording systematically influence how we perceive performances on record?

On the one hand, historical recordings are an amazingly rich and seemingly objective source of evidence about how music sounded in the past. Although listeners will readily recognize limitations of acoustic recordings from the early 20th century, these limitations decreased with improvements of recording techniques, and, even within these limitations, considerable information about the recorded music is preserved. This concerns, for example, relative variations in tempo and vibrato.

However, on the other hand, evaluations of historical recordings may be rather subjective. Most contemporary listeners are not familiar with the performing styles of the early 20th century, or are most of them familiar with the conditions of early recordings and the quality of the reproduction of early records. Moreover, the quality of historical recordings and the reproduction of these recordings may influence the perception of the recorded performances. For example, the limited frequency range of the recordings may limit the perception of consonants and timbre differences. Similarly the noisiness of reproductions may influence perception of volume or quality.

Whether listeners are able to listen through the differences in recording and reproduction quality and whether listeners are able to understand the intentions of performers even if the performing style is unfamiliar is unclear. So is the effect of this unfamiliarity on the perception of recorded performances. There is some evidence that the understanding of expressive intentions in performed music can be cross-

[a)]Current affiliation: "Music, Mind, Machine" group, Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands.
Electronic mail: r.timmers@nici.ru.nl or reneetimmers@solcon.nl

cultural (Balkwill and Thompson, 1999), which suggests that understanding is independent of familiarity with a performing style. On the other hand, other studies have shown an effect of musical training on perception and interpretation of performance (Repp, 1995; Honing, 2007; Timmers *et al.*, 2006), which instead suggests a dependence on familiarity with performing styles.

The reported study set out to examine the influence of the age of a recording and the quality of reproduction on the perception of recorded performances and to compare this to the influence of performance characteristics. The aim was to investigate this in the context of existing recorded material.

Four exploratory experiments were run that each considered a different aspect of perception of performance. The first experiment concerned the perception of the age of a recording, the second concerned the evaluation of the quality and emotionality of a performance, the third concerned the perception of emotional activity and valence, and the fourth concerned the perception of dynamics (see the following for further explanation and see the Appendix for the instruction of each experiment). Perception was assessed through subjective judgments on a rating scale.

All experiments used the same material: Four fragments from *Die junge Nonne*, a late song by Franz Schubert, sung by six famous sopranos reproduced in a "clean" and "noisy" version. The first recording is from 1907 and the latest from 1977. The four fragments consist of musical passages of the song with distinct emotional characteristics: The first and second fragments are negative in emotion in comparing earthly life with a roaring storm and the darkness of one's heart with the grave, while the third and fourth fragments are more positive in character: The nun finds peace in joining the convent. Additionally, the first and third fragments have high emotional activity, while the second and fourth fragments have relatively low emotional activity; the mood turns from distress (F1), depression (F2), and excitation (F3) to resignation (F4). This is the surface meaning of the text. Alternative interpretations include, for example, that finding peace through an "eternal marriage with God" is actually a metaphor for an escape from the torments of earthly life through death.

The two versions concerned clean and noisy reproductions of a recording. The original recording is the same, but the reproduction differs in noisiness: 78 recordings were transferred either in a "flat" way, i.e., without any processing, or they were cleaned using noise-reduction and anticlick software. Tape or digital recordings issued on CD are, on the other hand, already perfectly clean. To get two versions of these recordings, noise was added and the signal was low-pass filtered to some extent (details are explained in Sec. II).

Three analyses of the collected data were run addressing three specific sub-questions. The first analysis tested the effect of recording date and reproduction quality on the perception of performance. The aim of this analysis was to see to what extent subjective judgments of performances depend on recording date and noisiness of the reproduction. It tested whether our perception of performances is essentially influenced by conditions regarding the recording and the age of a performance or whether it is essentially independent.

The second analysis tested the effects of singer and fragment, and, most important, the interaction between these effects on perception of performance. The aim of this analysis was to see to what extent subjective judgments depend on a performer's interpretation of the music. Fragment alone may influence judgments, the overall style of a performer may influence judgments, and the specific interpretation of the music by a performer may influence judgments, resulting in an interaction between the effects of fragment and singer.

Finally, the third analysis tested the relationship between perception and measured aspects of the performances. Strong correlations between judgments and aspects of the performances provide suggestive evidence for the relevance of performance. The three analyses together should highlight the biasing effects of recording date and reproduction quality as well as the impact of characteristics of performances on perception judgments irrespective of recording conditions.

The rationales for the different experiments were the following. The judgments of the age of the recording were included to function as baseline for the ratings of the other experiments. It is a measure of how distant in time participants perceive the different recordings to be. It was the only measure that asked listeners to judge the recording, although they were advised to pay attention to both the recording and the performing style, since historical recordings can be cleaned, LP's can be noisy, and noise can be added to CDs. In all other experiments, listeners were explicitly asked to pay most attention to the performance.

The judgments of quality and affect (Exp2) are of interest, because they may highly depend on familiarity with performing style, as well as on recording/reproduction quality. Nevertheless, all recordings used in the study were of singers considered among the best of their time. It may be possible that participants do recognize the quality of past singers. Moreover, as observed by Day (2000), rhetorical and grand gestures in performance were stronger in early 20th century than in later 20th century performances. This may make earlier performances more emotionally affecting than later performances.

The judgments of perceived emotion (Exp3) are of interest, also because of an ambiguity in possible outcome: On the one hand, performing style changed considerably and therefore communication of a performer's intention to current day listeners may be difficult for older recordings. On the other hand, several authors and investigations have suggested that expression of emotion in singing and music performance has universal characteristics shared with expression of emotion in speech (Juslin and Laukka, 2003; Scherer, 1986; 1995; Sundberg, 1987). It would therefore be likely that communication of emotions is possible irrespective of recording date, as long as the relevant information is present. Indeed, analysis of the way singers express the different moods within Schubert songs showed high consistency between singers, over different time periods, despite evident changes in performing style (Timmers, 2007).

The judgments of perceived emotion were done using two rating scales: emotional valence and emotional activity. Valence and activity are two dimensions that distinguish well between different emotions (Russell, 1980). Emotions may

TABLE I. Overview of recordings used in the experiments.

Performers	Ref	Source
Susan Strong	SS 07	"Schubert Lieder on Record I, 1898-1939,"
Orchestra	Clean	EMI Classics 5 66150 2, 1997
Susan Strong	SS 07	HMV matrix 2004 f
Orchestra	Noisy	
Susan Metcalfe-Casals	SMC 37	"Schubert Lieder on Record II, 1929-1952,"
Gerald Moore	Clean	EMI Classics 5 66154 2, 1997
Susan Metcalfe-Casals	SMC 37	HMV matrix CTPX 3884-1
Gerald Moore	Noisy	
Lotte Lehmann	LL 41	"Lotte Lehmann: Schubert," LYS 231-234, 1997
Paul Ulanowsky	Clean	
Lotte Lehmann	LL 41	Columbia matrix XCO 30013-1
Paul Ulanowsky	Noisy	
Elisabeth Schwarzkopf	ES 52	"Schubert: 12 Lieder, 6 Moments musicaux,"
Edwin Fischer	Clean	EMI Classics 5 67494 2, 2000
	Noisy	
Elly Ameling	EA 75	"Schubert Liederen,"
Dalton Baldwin	Clean	Philips 464 334-2, 1999
	Noisy	
Gundula Janowitz	GJ 77	"Schubert Lieder,"
Irwin Gage	Clean	Deutsche Grammophon 453 082-2, n.d.
	Noisy	

have positive or negative valence, such as happiness compared to anger, and they may have high or low arousal, such as anger compared to sadness or depression. The use of these dimensions was preferred over the use of specific emotion words, because it allows for subtle distinctions between performers to come forward: Overall, a musical passage may be perceived to be negative. Within this overall tendency, one performance may be perceived to be more negative than another. These subtle differences are hard to express in words, and listeners tend to disagree on terminology when asked to characterize music in subcategories (Gabrielsson and Juslin, 2003).

Finally, the judgments of dynamics are of interest in two respects (Exp4). First, comparison between judgments of dynamics and measurements of amplitude is a test for the reliability of amplitude measurements. Second, it is of interest to compare the perceived range in dynamics for historical recordings with that of modern recordings. It is likely that historical recordings tend to have a smaller dynamic range than nowadays possible. Acoustical recordings were very noisy and needed a loud signal for a proper signal to noise ratio (Gelatt, 1956). On the other hand, an overload of the cutter due to too loud sounds had to be avoided as well. The situation improved with the introduction of microphones and amplification with electrical recording. Nevertheless, very soft and loud sounds remained problematic. Recording engineers started to control the recorded signal and often "tamed" the performed dynamic range to avoid overload and ensure audibility (Copeland, 1991).

It should be noted that this is an exploratory study that uses existing recorded material. This makes the study interesting for music research on recordings and ensures ecological validity. The drawback is, however, that the results are not entirely clear-cut: The effect of recording date on perception of performance combines the effect of performing style and recording conditions. The effect of singer similarly combines the effect of performer and recording conditions. Therefore, the three analyses are needed to come to a complete interpretation of the data.

II. METHOD

A. Musical material

Six performances of *Die junge Nonne* were selected from a database of recordings. *Die junge Nonne* is one of the songs by Schubert that has been recorded regularly throughout the 20th century. As mentioned earlier, characteristic of the song is that it contains a succession of moods.

The aim was to have a set of early performances, in a relatively unknown style, and a set of later performances in a more familiar style. This aim was counterbalanced with the aim of having performances spreading a time period more evenly. The result was the choice for three performances from before 1945, and three performances from after 1950, assuming a break in performing style around the second world war as was observed by Philip (1992). The restriction to six performances in total was made to limit the total number of stimuli to be used in the experiments. Details of the recordings used in the study are listed in Table I.

Four fragments from each performance were selected to serve as musical material. Each fragment has a specific mood: Fragment 1 (F1, bars 36–41) is high in activity and negative in mood. Fragment 2 (F2, bars 43–49) is low in activity and negative in mood. Fragment 3 (F3, bars 54–61) is high in activity and positive in mood, while Fragment 4 (F4, bars 71–74) is low in activity and positive in mood. The moods of each fragment were determined in a previous study (Timmers, 2007), based on the meaning of the text as well as structural aspects of the composition.

Two versions were used of each selected performance—one clean and the other noisy. The clean versions were taken from commercially issued CDs. To acquire an even cleaner version of the recording of Elisabeth Schwarzkopf, the few cracks and clicks in the digitized recording were diminished using declicker and denoiser functions of an audio editing program.

In this context, clean and noisy should not be interpreted in absolute terms, but relatively: Clean means relatively clean compared to its respective noisy version, and noisy means relatively noisy. A clean 78 is not as free from noise and clicks as a modern recording. Likewise, a clean acoustical 78 recording has a lower signal to noise ratio and is more limited in frequency range than a clean electrical 78 recording. The main point of having a clean and noisy version of a recording is to have two versions that differ in transfer of the original performance and that can be considered different in quality of reproduction.

The noisy versions of the 78 recordings were acquired by making a flat transfer of the recording: The original 78s were played back using a modern turntable. The analogue output from the turntable was led to an amplifier and into an analog-to-digital converter. The digital output was led into a personal computer and was recorded.

To obtain a noisy version of the recordings of Elly Ameling and Gundula Janowitz, noise had to be added to the recordings. Additionally, the recordings were modified to sound "older." First, the signal was compressed. Second, noise was added. Noise was acquired by recording the playback of a blank SP shell ac disc from 1950, which was used to add noise to the recording of Elly Ameling, and a blank SP shellac disc from 1935, which was used to add noise to the recording of Gundula Janowitz. Finally, the mixed audio track was bandpass filtered by reducing the amplitude gradually below 50 Hz (monotonically toward 30 Hz) and above 3000 Hz (monotonically toward 5000 Hz). This means that also the noise was deamplified at higher and lower frequencies. This enhanced the integration between the noise and the signal of the recorded performance. As a final modification, all audio files were saved as mono tracks and the resolution was set to 22000 Hz 16 bits.

B. Participants

All participants had more than ten years of formal musical training. Most of them were university music students (second year and higher). The others were advanced performers. Participants had a variety of nationalities (e.g., British, Dutch, American, Israeli, Greek, Japanese). Two were German native speakers. Most Dutch participants were able to understand German. All participants had a background in classical music.

Participants did three experiments in a row to limit the time per participant. Participants were randomly assigned to the experiments. In total, the number of participants was 22 for Exp1, 26 for Exp2, 40 for Exp3, and 32 for Exp4. The main reasons for the number of participants to vary over experiments were practical. Originally, two additional experiments were run.

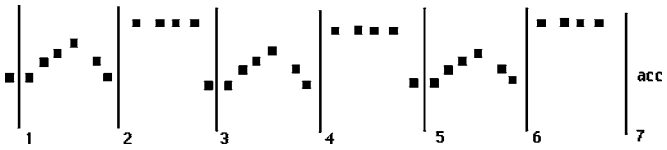

FIG. 1. Answer sheet for Fragment 1 of Exp4. The dots represent the melody of the first fragment. Bars are indicated by vertical lines and numbers.

C. General procedure

Participants were seated behind a laptop and read the instructions from a print out (instructions are given in full in the Appendix). After a general introduction to the experiments, the instruction was given for the first experiment. The participants started the experiment immediately without a practice trial. They put on headphones (Sony MDR-7506) and used a mouse to play a stimulus, to give the ratings, and to press the ok/save button to go to the next stimulus (all programmed in POCO, see Honing, 1990). Sound levels of the playback were set to a comfortable level and fixed throughout the experiments. The labels above the radio buttons of the rating scales changed with experiment. Either one or two rating scales were used depending on the experiment. The order of the two rating scales was counterbalanced between participants. The presentation order of the musical stimuli was randomized over participants.

Separate answer sheets were used for Exp4. For this experiment, the computer interface was only used to play stimuli. The answer sheets showed a representation of the sung melody of a musical fragment and bar lines were indicated. This was necessary, because the participants indicated the dynamics per bar. Figure 1 shows the representation of the melody of F1.

Because the stimuli were presented randomly to the participants, participants did not know beforehand which fragment would sound. For assistance, the fragment number was indicated before each stimulus in Exp4 using a computer voice mentioning the fragment number.

As mentioned before, all participants did three experiments in a row. The instruction for each experiment was given just before the start of an experiment. The instructions are given in the Appendix. For Exps 1, 2, and 3, all 48 stimuli were presented in random order after each other. These 48 stimuli included six performances of four fragments in two versions — clean and noisy. For Exp4, the 48 stimuli were split in half: half of the participants judged the noisy versions of the recordings and the other half judged the clean versions of the recordings. This was done to restrict the duration of this experiment. Each experiment took approximately 20 min, which resulted in an overall duration of about an hour per participant.

III. RESULTS

A. Effect of date and version

First, the effects of recording date and version on the judgments were examined. Average ratings over fragment were used for this analysis. Figure 2 shows the mean ratings per recording date and version for the variables of the differ-

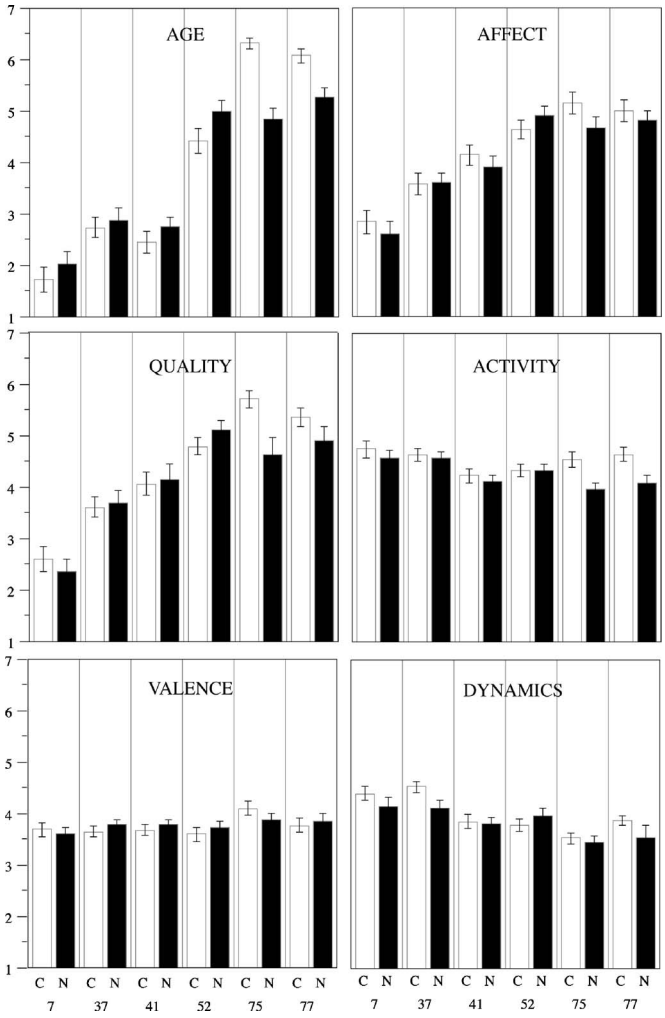

FIG. 2. Average and standard errors of ratings of Exps 1–4 per date and version.

ent experiments. It can be seen that the slope of the relationship between recording date and judgment is steep for age, and quality and almost flat for valence and activity. The effect of version is small for all judgments. It is generally larger for modern recordings that were made noisy than for older recordings, especially for the judgments of age, quality, and activity.

Table II reports the explained variances and significance values of the results of a series of mixed model ANOVAs. Each ANOVA had date (continuous variable) and version (nominal variable) as independent variables and one of the judgements as dependent variable. In a mixed model ANOVA, participants are treated as random effect and date and version as fixed effects. Date and version are within subject effects for all experiments except for Exp4, which varied version across participants. The recommended residual maximum likelihood method was used as estimation method.

The results confirm the observations made with respect to Fig. 2; the effects of date were considerably stronger than the effects of version or the interaction between date and version. The effect of date was small for judgments of activity and valence.

To examine, in addition, whether the range in responses changed systematically over time, the analyses were rerun using the standard deviation of responses over fragments within singers as data points. The hypothesis was that, for some judgments, the variation in judgments could be smaller for early singers than later singers. For example, the variation in dynamics was predicted to be more restricted for early recordings than for later recordings, because of an expected smaller range in volume differences between musical fragments.

The effect of date was significant, but small, for all judgments, except quality. The effect of date was relatively strong for judgments of dynamics ($p<0.0001$) and valence ($p<0.01$). Focusing on these stronger effects, the amount of variation increased over time suggesting restricted variation in perceived dynamics and valence for earlier performances (see Fig. 3). The effect of version was significant for judgments of quality only ($p<0.001$). Variation in perceived quality was more restrained for the noisy versions (all tended to be lower in quality) than the clean versions.

B. Effect of fragment and singer

The second analysis of the data examined the effect of fragment and singer and specifically the interaction between these effects on the judgments: Did singers communicate a personal interpretation of the musical fragments?

TABLE II. Summary of results of mixed model ANOVAs for Exp1 ($N=22$), Exp2 ($N=26$), Exp3 ($N=40$), and Exp4 ($N=20$) testing the effects of recording date (one level) and version (two levels). Partial explained variances (R^2), F ratios, and p values are given for significant effects.

	Date			Version			Date ×Version[a]		
	R^2	F	p	R^2	F	p	R^2	F	p
Exp1									
Age	0.60	472	<0.0001			n.s.	0.02	17.1	<0.0001
Exp2									
Affect	0.35	186	<0.0001			n.s.			n.s.
Quality	0.38	270	<0.0001			n.s.	0.01	4.02	<0.05
Exp3									
Activity	0.02	13.1	<0.0001	0.02	15.7	<0.01	0.01	6.36	<0.05
Valence	0.01	9.29	<0.01					n.s.	n.s.
Exp4									
Dyn.	0.23	76.5	<0.0001			n.s.			n.s.

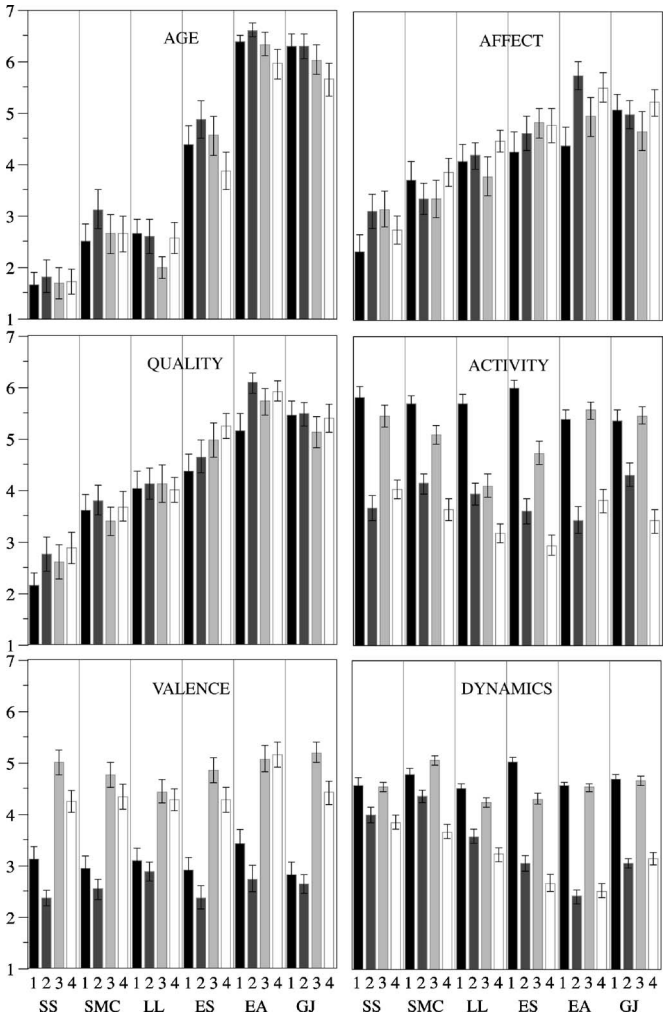

FIG. 3. Average and standard errors of ratings of Exps 1–4 per singer and fragment of the clean versions of the recordings. Singers are ordered according to recording date.

For this analysis, only the judgments of the clean versions of the recordings were used. A series of repeated measures ANOVAs were used to test the effects of fragment (nominal) and singer (nominal) and the interaction between them for each judgment separately. For Exp4, data consisted of the average rated dynamics per performance (averaged over bars).

Figure 3 and Table III show summaries of the results. For most of the judgments, there is only one variable that was significant or highly significant and contributed most to the explained variance. This is singer for the judgments of age, affect, and quality, and fragment for the judgments valence and activity. For dynamics, both effects of fragment and singer are highly significant. The interaction between fragment and singer is highly significant for the judgments of activity and dynamics. It is just significant for the judgments of valence.

These results confirm the division observed in the first analysis between the judgments of age, quality, and affect, on the one hand, and the judgments of emotional activity, valence, and dynamics, on the other hand. The first group of judgments vary strongly with recording date. They also intercorrelate strongly: The correlation is 0.86 for average perceived age and affect, 0.92 for average perceived age and quality, and 0.96 for average perceived affect and quality. The judgments of the second group depend, however, more strongly on the musical fragment and, especially for perceived emotional activity and dynamics, the performers' interpretation of the music. The judgments of dynamics and activity are strongly correlated (r=0.84).

C. Correlations with characteristics of the performances

The final analysis of the data examined the relationship between judgments and aspects of the performances. Ratings averaged over participants of the clean versions of the performances were correlated with measurements of the performances. These measurements were made in a previous study (Timmers, 2007). The measurements included duration of each bar in seconds, the average sound level of each bar, the average vibrato rate of a long note in each bar in cycles per second, the extent of a large vibrato cycle of a long note in

TABLE III. Summary of results of repeated measures ANOVA for Exp1 (N=22), Exp2 (N=26), Exp3 (N=40), and Exp4 (N=22) testing the effects of fragment (four levels) and singer (six levels) on the judgments of the clean recordings. Partial explained variances (R^2), F ratios, and p values are given for significant effects.[a]

	Fragment			Singer			Fragment × Singer		
	R^2	F	p	R^2	F	p	R^2	F	p
Exp1									
Age	0.01	3.46	<0.05	0.62	105	<0.0001			n.s.
Exp2									
Affect			n.s.	0.22	21.3	<0.0001			n.s.
Quality			n.s.	0.40	47.9	<0.0001			n.s.
Exp3									
Activity	0.33	88.4	<0.0001	0.02	3.43	<0.05	0.05	6.26	<0.0001
Valence	0.29	41.6	<0.0001	0.01	3.14	<0.05	0.01	2.02	<0.05
Exp4									
Dyn.	0.52	184	<0.0001	0.11	41.0	<0.0001	0.10	15.7	<0.0001

[a]Effects are significant using Greenhouse-Geisser epsilon for violations of sphericity for effects with larger number of levels.

TABLE IV. Significant correlations ($p<0.05$) between mean judgments (rows) and measurements (columns) of the clean recording of each fragment and each singer.

	Sound level	Bar duration	Vibrato extent	Vibrato rate	Up	Down
Age	−0.78	0.41		−0.62		
Affect	−0.89	0.62		−0.71		
Quality	−0.89	0.63		−0.75		
Activity			0.70		0.53	−0.41
Valence		0.50				0.45
Dynamics	0.63		0.57	0.44	0.55	

each bar in semitones, and the number of pitch glides up and down in each bar. Table IV shows the significant correlations.

All judgments show significant correlations with several aspects of the performances. Many of the correlations are high, except for valence, which shows only moderate correlations with performance aspects.

To interpret these correlations, it is useful to take the results of the measurement study (Timmers, 2007) into account. From the measurements, several systematic changes in performing style within the 20th century were observed. The amount of rubato tended to decrease over time, global tempi tended to decrease, later performances tended to be softer on average than early performances, vibrato rate decreased gradually over time, while vibrato extent increased over time, and the number of pitch glides was medium in the beginning of the 20th century, increased toward the 1930s, and decreased after the 1940s.

The correlations reported in Table IV partly reflect these changes over time. Judgments of age correlate negatively with average amplitude and vibrato rate and positively with average bar duration. Judgments of quality and affect also show these correlations. Notably, the correlations with quality are highest, suggesting that quality is strongly related to performing style.

Dynamics and emotional activity are, on the other hand, correlated with vibrato extent and number of pitch glides up or down. Perceived dynamics is correlated with measured amplitude, but the correlation is not very high. This suggests only limited reliability of the measurements of sound level to represent dynamics.

Significant correlations with perceived valence include medium correlations with average bar duration and number of downward pitch glides. These correlations confirm the relationship between valence and aspects of performances for this particular song as observed in Timmers (2007): Positive passages tended to be slower in tempo and had more downward pitch glides than negative passages. This can be understood if we interpret the positive passages to be a release of the negative tension rather than, e.g., a positive excitation or uplift.

Perceived dynamics was the only judgments that participants rated per bar. For this judgment, one more analysis was done and judgments per bar were correlated with measurements of sound level per bar for each singer individually. Table V shows the correlations that were significant. Notably, the correlations are now considerably higher than in Table IV for all modern recordings, starting from the recording of Lotte Lehman from 1941. The older recordings show lower correlations with an insignificant correlation between measured sound level and perceived dynamics for the recording from 1937.

A possible reason for the insignificance of the correlation for the recording of Susan Metcalfe-Casals from 1937 is the great difference in amplitude between voice and piano. If the singer sings only half a measure, the measured sound level drops considerably, while participants may rate the dynamics as forte based on the dynamics of the voice. Perceived loudness can be corrected for presence or absence of the voice by multiplying the judged dynamics with the fraction of the bar that the singer sings. After this rough correction of the judged dynamics, the correlation between measured sound level and perceived dynamics is significant ($r=0.60$, $p<0.01$).

The discrepancy between the correlation between perceived dynamics and measured sound level reported in Table IV and the correlations reported in Table V suggests that measurements of sound level capture relative variations in dynamics within a given recording more reliably than the relative loudness of different recordings. This is not necessarily a deficit of the measuring method, but may also be due to subjective perception of dynamics and the task of the participants in Exp4. The participants were instructed to write down the variations in dynamics within a musical fragment and used the scale for this. Their task was not to compare the

TABLE V. Significant correlations between mean judgments of dynamics per bar and measurements of sound level per bar, calculated for each singer separately.

	SS 07	SMC 37	LL 41	ES 52	EA 75	GJ 77
Sound level	0.52		0.83	0.88	0.92	0.88

relative loudness of different recordings. The indication of relative loudness of different recordings was only an indirect result of the task to indicate the dynamics within a performance.

IV. DISCUSSION AND CONCLUSION

The aim of the study was to examine the influence of the age of a recording and the quality of reproduction on the perception of recorded performances and to compare this to the influence of performance characteristics. This was done in an exploratory study using commercial historical and modern recordings.

Clear tendencies of judgments to systematically change with recording date were observed for all perceptual aspects. Judgments of age and quality changed most strongly with recording date followed by judgments of affect and dynamics. Judgments of perceived emotion were most independent of recording date. Additionally, the variation in communicated emotional valence and dynamics over musical fragments tended to be more restrained for older recordings than for modern recordings.

Tendencies of judgments to change with recording version were, in contrast, small for all judgments and significant for only a few perceptual aspects. This suggests that listeners were able to abstract relevant information from specific reproduction conditions.

The importance of performance characteristics was suggested by a significant interaction between the effects of fragment and singer for the judgments of perceived emotion and dynamics. This interaction highlights the influence of the performer on the perception of the musical fragments, which suggests communication of a personal interpretation of the music.

Finally, significant correlations between measured characteristics of the performances and perceptual judgments were observed for all judgments. This suggests that not only the judgments that showed an interaction between fragment and singer, but also the judgments that varied most strongly with recording date may have varied due to changes in performance characteristics.

In short, the main result of the study was the clear division between perceptual judgments that varied strongly with recording date and perceptual judgments that varied less strongly with recording date. Additionally, a limited effect of reproduction version was observed and strong correlations between perceptual judgments and measured performance variables, suggesting that the actual influence of the recording is limited compared to the influence of performance characteristics.

However, it should be noted that this conclusion is drawn tentatively. The study used existing recorded material. While this enhanced ecological validity, it limited the control over the experimental material. The effects of version and date are not single effects, but consist of different variables: Recording date implies differences in performing style as well as recording conditions and reproduction conditions, while reproduction version consists of differences in noisiness, frequency range, and possibly other aspects such as dynamic compression or boost. Each experiment could be refined and specific effects examined. Additionally, it might be useful to define sensitivity thresholds: For example, noisiness may influence perceptual judgments from a specific noise level onwards. The specific level may vary with perceptual aspect.

Nevertheless, the study generated interesting results, adding to a growing literature on perception of performance. For example, it showed a strong association between judgments of quality and affect, which emphasizes the possible importance of aesthetic experiences for emotional affect (e.g., Scherer, 2004). In contrast, emotional affect was not strongly related to judgments of emotional activity, which suggests that felt emotional impact is quite different from perceived emotional activity (Gabrielsson, 2001). Emotional affect was also negatively correlated with sound level, while emotional activity correlated positively with sound level. This further emphasizes the complexity of stimulus-response relationships for emotional arousal.

The high correlations observed between variations in judged dynamics and measured sound levels per bar are promising for research that uses measurements to assess performance characteristics. Nevertheless, the exact relationship between perception and measurement needs to be further examined in future research. Part of the complexity of perception of dynamics was highlighted by showing a possible focus of listeners on the voice when judging dynamics.

Part of the contribution of the study is to raise issues for further research. In being explorative, it addressed perception of recorded performances rather broadly. It distinguished between perceptual judgments that are highly sensitive to differences in recording date and judgments that are almost independent of recording date. It remains an interesting issue how listeners perceive old recordings. Listeners seem certainly able to listen through differences in manners of record reproduction and judgments are strongly associated with performance variables. Nevertheless, standards have clearly changed, which influence judgments of quality and age. Additionally, changes restrict to some extent the communication between early recorded performers and modern listeners.

ACKNOWLEDGMENTS

This research was funded by the Arts and Humanities Research Council's Research Centre for the History and Analysis of Recorded Music (CHARM). I would like to thank Daniel Leech-Wilkinson for his help during the preparation of the studies and for collecting the Schubert material. My thanks also go to Karsten Lehl for the flat transfers of the early 78 recordings, and to Roger Beardsley for the recordings of blank shellac 78s.

APPENDIX

1. Instruction experiment 1

In this experiment, you will hear short fragments of recordings of Schubert songs. Your task is to decide for each fragment if you think it is an historical recording from before 1945 or a modern recording from after 1950. You do this on

a scale from 1 to 7. 1 stands for certainly before 1945, and 7 for certainly after 1950. Try to use levels other than 4, and the extremes, as much as possible. Please note that the amount of noise or clicks is not a good criterion for the "age" of a recording, since historical recordings can be cleaned (noise is filtered out) and LP recordings can also have cracks and noise, and noise can be added digitally to modern recordings. Therefore pay most attention to the performance and try to base your answer on that. There will be 48 fragments in total.

2. Instruction experiment 2

In this experiment, you will hear short fragments of recordings of Schubert songs. Your task is to evaluate the quality of the performance and how much the performance affects you emotionally on a scale from 1 to 7. Please try to use the entire scale — both the extremes and the middle levels.

The reason to use these two rating scales is that you may consider a performance to be "good" and "well-performed" (the quality is high), but at the same time the performance may not affect you emotionally (affect is low). Other performances may be less "perfect" (quality is low), but may touch you much more (affect is high). There will be 48 fragments in total.

3. Instruction experiment 3

In this experiment, your task is to indicate the emotion you perceive "in" the performance. You do this by characterizing the perceived emotion along two dimensions—valence and activity. The term valence is used to indicate whether the perceived emotion is positive or negative. The term activity is used to indicate whether the perceived emotion is active or passive.

The dimensions of valence and activity were found by different researchers to give a suitable summary of emotions and relations between them. Some emotions have activity associated with them, such as joy and anger, while other emotions have passivity associated with them, such as sadness, and boredom. In addition, some emotions are seen as positive, while others are considered negative.

In this experiment, you indicate valence and activity on a scale from 1 to 7 for 48 fragments of recorded performances of Schubert songs. When rating valence, 1 stands for negative and 7 for positive. When rating activity, 1 stands for low and 7 for high. Please try to use the entire scale, so try to use both the extremes as well as the middle levels. Note that the emotion you feel may be different from the emotion you perceive in the performance. In the current task, we are interested in the communication from performer to listener. So we would like to know what intended emotion you perceive rather than how much the music affects you.

4. Instruction experiment 4

In this experiment, you notate the dynamics of a performance and you do this for 24 fragments of recorded performances of Schubert songs. Listen to the music and start notating the dynamics on the sheet, below the representation of the respective melodic line. Before each fragment, you will be told which of the four musical excerpts will sound. Indicate levels of dynamics using pp to ff (or 1–6 if you find that easier). Indicate changes in dynamics also by using pp to ff and not by writing crescendo or decrescendo. Please make one marking of dynamics per bar. Note that different musicians perform the same musical excerpts. We are interested in differences between performers — so even if they perform the same music, you may perceive that they use different dynamic levels. You may listen to a performance more than once if necessary.

Balkwill, L. L., and Thompson, W. F. (**1999**). "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," Music Percept. **17**, 43–64.

Copeland, P. (**1991**). *Sound Recordings* (British Library Board, London).

Day, T. (**2000**). *A Century of Recorded Music: Listening to Musical History* (Yale University Press, New Haven).

Gabrielsson, A. (**2001**). "Perceived emotion or felt emotion: Same or different?," Musicae Scientiae, Special issue, 2001/2002, 123–147.

Gabrielsson, A., and Juslin, P. N. (**2003**). "Emotional expression in music," in *Handbook of Affective Sciences* edited by R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Oxford University Press, Oxford, pp. 503–534).

Gelatt, R. (**1956**). *The Fabulous Phonograph: The Story of the Gramophone from Tin Foil to High Fidelity* (Cassell & Company, London).

Honing, H. (**1990**). "POCO: An environment for analysing, modifying and generating expression in music," in *Proceedings of the 1990 International Computer Music Conference* (CMA, San Francisco), pp. 364–368.

Honing, H. (**2007**). "Is expressive timing relationally invariant under tempo transformation?," Psychol. Music. **35**, 1–10.

Juslin, P. N., and Laukka, P. (**2003**). "Communication of emotions in vocal expression and music performance: Different channels, same code?," Phys. Bl. **129**, 770–814.

Philip, R. (**1992**). *Early Recordings and Musical Style: Changing Tastes in Instrumental Performance, 1900-1950* (Cambridge University Press, Cambridge).

Repp, B. H. (**1995**). "Detectability of duration and intensity increments in melody tones: A partial connection between music perception and performance," Percept. Psychophys. **57**, 1217–1232.

Russell, J. A. (**1980**). "A circumplex model of affect," J. Pers Soc. Psychol. **39**, 1161–1178.

Scherer, K. R. (**1986**). "Vocal affect expression: A review and a model for future research," Polymer **99**, 143–165.

Scherer, K. R. (**1995**). "Expression of emotion in voice and music," J. Voice **9**, 235–248.

Scherer, K. R. (**2004**). "Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?," J. New Music Res. **33**, 239–251.

Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois University Press, DeKalb).

Timmers, R., Marolt, M., Camurri, A., and Volpe, G. (**2006**). "Listeners' emotional engagement with performances of a Scriabin étude: An explorative case study," Psychol. Music **44**, 481–510.

Timmers, R. (**2007**). "Vocal expression in performances of Schubert songs on record." Musicae Scientiae **XI**, 73-101.