ARCHIVES OF ACOUSTICS
2, 4, 257 (1977)

VIBRATO AND VOWEL IDENTIFICATION

J. SUNDBERG

Royal Institute of Technology, S-100 44 Stockholm 70 Center for Speech Communication Research and Musical Acoustics

The influence of vibrato on vowel identification is studied in synthesized vowel sounds with fundamental frequencies between 300 and 1000 Hz. Phonetically trained subjects were asked to identify these stimuli presented with and without vibrato as any of 12 Swedish long vowels. Small and occasional effects are observed on the mean formant frequencies and the scatter of the responses.

However, the effects are greater and more frequent than could be expected to occur by chance. In the majority of the cases, where the vibrato affected the responses, the vowel identification became somewhat harder when the stimulus was presented with vibrato. It is assumed that the vibrato may be used to conceal some of the deviations from the vowel qualities in speech which are typically made in singing.

1. Introduction

Vibrato occurs in music performed on many types of instruments, such as string and wind instruments, and in singing. Physically, it corresponds to a quasi sinusoidal modulation of the fundamental frequency. Provided that the rate and extent of the vibrato is kept within certain limits, the vibrato is generally considered as useful means of musical expression in a variety of musical styles.

Acoustic characteristics typical of music sounds can be assumed to serve some purpose in musical communication. It is frequently assumed that the vibrato covers up small errors in the fundamental frequency. In a previous investigation no support was found for this hypothesis as far as the pitch of single tones – and thus not chords – is concerned (SUNDBERG 1972). Here, another frequently assumed explanation of the vibrato will be tested, namely that the vibrato facilitates vowel identification. This hypothesis seems plausible, because as the partials move in frequency (e.g. owing to a vibrato) their amplitudes scan small parts of the vocal-tract transfer function, and more

information concerning the formant frequencies is communicated. The effect would be particularly strong in the soprano range where the fundamental frequency may be even higher than 1000 Hz. In such cases the information on the formant frequencies would be hard to decode from the acoustic spectrum.

2. Experiment

A series of high-pitched sounds was synthesized with and without vibrato using a terminal analogue. Six formant-frequency combinations were used, each of which corresponds to a Swedish long vowel: [u, o, ɑ, e, i, y]. The vowel formant frequencies (see Table 1) were found in a soprano singing at a funda-

Table 1. Formant frequency values (in Hz) used for the synthesis of vowel stimuli and ascribed to response vowels

Vowel	Vowel stimuli				Response vowels		
	F_1	F_2	F_3	F_4	F_1	F_2	F_3
u	325	700	2700	4000	320	565	2775
o	430	650	2820	4000	415	725	2775
a	640	1090	2800	4000	700	1065	3000
ɑ					855	1200	2820
æ					815	2030	3000
ɛ					625	2305	3040
e	400	2000	2700	4000	375	2790	3360
i	270	1900	1950	4000	275	2675	3655
y	250	1950	2510	4000	275	2555	3210
ʉ					300	1920	2745
ø					390	2040	2725
œ					565	1290	2690

mental frequency of 262 Hz (SUNDBERG 1975). Contrary to the effect which had been observed in this soprano, each of these 6 combinations of formant frequencies was used «without changes» for 4 different pitches within the soprano range: 300, 450, 675, and 1000 Hz. In this way the material included sounds ranging from maximum ease to maximum difficulty as regards vowel identification. The vibrato consisted of a sinusoidal modulation of the fundamental frequency. The modulation rate was 6 undulations per second, and the deviation was ±50 cents. These values are found in normal singing, even though the deviation is normally sligthly narrower.

Four samples of each vowel stimulus were tape recorded at a constant overall amplitude. The time and the shape of the onsets and decays were all similar. The stimuli samples were arranged in random order avoiding more than one repetition of a given stimulus in sequence. The entire tape contained a total of 192 vowel samples: 6 formant-frequency combinations × 4 fundamental

frequencies $\times 4$ presentations $\times$ 2 cases (with and without vibrato). Each stimulus had a duration of 1 s and was followed by a silent interval of 4 s.

The stimuli were presented through head-phones (Sennheiser HD 414) at an overall *SPL* of 70 dB, approximately, to 10 phonetically trained subjects. Their task was to decide which one of 12 given Swedish long vowels [u, o, ɑ, a, æ, ɛ, e, i, y, ʉ, ø, œ] sounded most similar to the stimulus they heard and to write that vowel in phonetic symbols on a test chart. These vowels, suggested by the subjects as interpretations of the stimuli, will be referred to as «response vowels». The test, including 3 short breaks, lasted about 30 min.

3. Evaluation of responses

It seems reasonable to hypothesize that the process of identifying an isolated stimulus as a specific vowel involves three major steps. The first is a purely sensory one in which the acoustic stimulus is converted into some kind of «sensory information». The second step involves the extraction of an ensemble of «sensory characteristics» from this sensory information. In the third step a «comparison» is made between these sensory characteristics and corresponding characteristics of vowels stored in the subject's internal reference. The difficulty in identifying a sound as a specific vowel would depend on the degree of similarity between the ensemble of sensory characteristics and the characteristics in the internal reference.

The vibrato may affect the sensory information in various ways. Let us assume that it affects the extraction of sensory characteristics. If so, the responses for a given vibrato stimulus should differ from those obtained when the same stimulus was presented without vibrato. In the test responses the effect would then be that the mean formant frequencies of the response vowels for a given stimulus would be different in the cases with and without vibrato. In that case the identification difficulties may be affected in various ways, depending on whether the vibrato makes the sensory characteristics (1) more similar or (2) less similar to the subject's internal reference. The cases (1) and (2) would differently affect the scatter of the responses.

The above considerations demonstrate the need for some measures of the average and of the scatter of the response vowels. In order to obtain such measures a set of formant frequencies was ascribed to each of the 12 response vowels in accordance with data on female speakers (FANT 1973). Thereby the fourth formant, having negligeable influence on the perceived vowel identity, was disregarded. Those values are listed in Table 1. These formant frequency values were then expressed in Mels and for each stimulus the averages and the standard deviations of each of the three lowest formant frequencies of the response vowels obtained from that stimulus were computed. The resulting set of three mean formant frequencies was considered to represent the «average response vowel» of that stimulus.

The scatter between the responses obtained from a given stimulus was estimated in the following way. Recently, LINDBLOM (1975), following ideas of PLOMP (1970), established that the perceptual distance D between any two Swedish vowels can be approximated as

$$D = \sqrt{\Delta M_1^2 + \Delta M_2^2 + \Delta M_3^2},$$

where ΔM_n is the frequency difference in Mels in the nth formant between the two vowels. This equation was used to calculate for each stimulus the perceptual distance between the average response vowel and each of the individual responses to the same stimulus. Then, the average of these distances over subjects was determined for each stimulus. The resulting values were assumed to reflect the «scatter of the responses», i.e. the difficulty with which the stimulus was identified as a specific vowel.

4. Results

One of the subjects refused to give any interpretation of the stimuli, all of which happened to lack vibrato. The inter-subject differences regarding the internal reference appeared to be rather small.

The formant frequencies of the average responses are shown in Fig. 1, and the corresponding standard deviations are listed in Table 2. The formant frequencies, used in the synthesis of the stimuli, are also indicated in Fig. 1. Hence, the relationships between these formant frequencies and those of the average responses can be examined. The stimuli synthesized with [ɑ]-formants offer an illustrative example. The formant frequencies of the average response drop with rising fundamental frequency up to 675 Hz. The reason for this is probably that, in order to maintain the vowel quality when the fundamental is raised, it is necessary to raise the formant frequencies slightly. SLAWSON (1968) found an average increase of 10 % in the formant frequencies to be required per octave rise of the fundamental.

If the fundamental is raised while the formant frequencies are kept constant, as was done in our experiment, the vowel quality will shift towards a vowel with lower formant frequencies. In this case the subjects probably identified the two lowest formant frequencies correctly, and the formant frequencies ascribed to the response vowels contain an error which depends on the fundamental frequency. A correction would be needed, but it is not known exactly how to make this correction. For instance, SLAWSON's correction of 10 % increase in formant frequency per octave rise in the fundamental does not fully compensate the effect. It should also be mentioned that all stimuli with a fundamental frequency of 675 Hz gave almost invariously response vowels in which one of the (assumed) formants matched one of the stimulus partials in frequency almost perfectly.

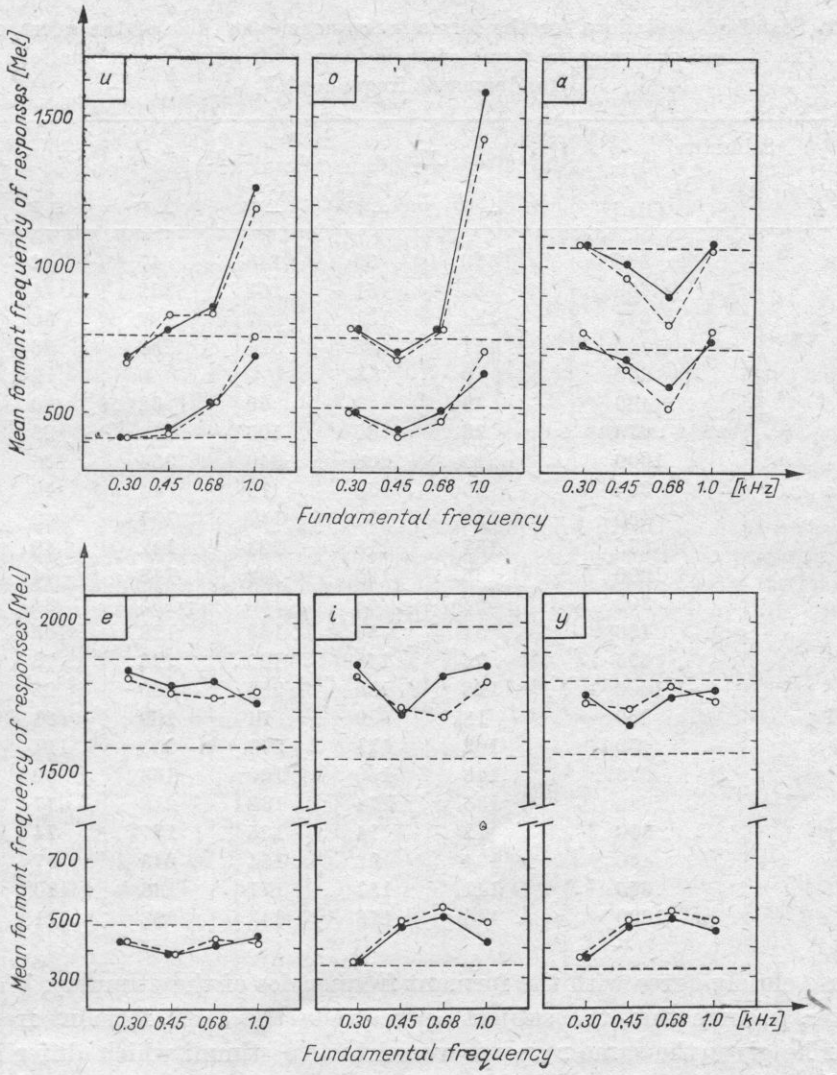

Fig. 1. Evaluation of the response vowels in terms of the means of their assumed two lowest formant frequencies expressed in Mels. Filled and open circles pertain to stimuli presented with and without vibrato, respectively. Horizontal dashed lines show the formant frequencies which were used in the synthesis of the stimuli

This may indicate that, in very high-pitched vowels, subjects tend to interpret the frequency of a spectrum partial as the frequency of a formant. This would lend support to the assumption that the formant frequencies which we ascribed to the response vowels may agree with the formant frequencies «perceived» by the subjects in very high-pitched vowels. Additional support for the same speculation may be found in the case of the [ɑ]-formant stimulus at 1000 Hz fundamental frequency. Here, the formant frequencies of the average

Table 2. Standard deviations for the formant frequencies in Mels of the average response vowels. The stimuli are given in terms of their formant-frequency combinations (F_N) and fundamental frequency (F_0)

Stimulus		M_1		M_2		M_3	
F_N	F_0 [Hz]	$-V$	$+V$	$-V$	$+V$	$-V$	$+V$
u	300	30	33	146	47	2	0
	450	59	57	251	351	6	42
	675	134	131	227	187	36	36
	1000	167	145	324	290	86	53
o	300	0	0	0	0	0	0
	450	49	43	69	62	0	0
	675	96	66	105	158	26	14
	1000	213	206	340	354	106	99
ɑ	300	60	55	73	44	46	33
	450	124	130	147	154	49	47
	675	151	118	203	142	40	31
	1000	135	125	202	148	44	35
e	300	65	65	150	155	84	94
	450	51	50	133	128	105	83
	675	96	141	120	226	76	101
	1000	125	139	296	236	95	102
i	300	18	9	74	109	81	110
	450	102	121	273	208	111	101
	675	146	150	120	163	80	84
	1000	135	174	132	216	77	106
y	300	12	14	125	137	74	81
	450	104	92	154	243	97	114
	675	141	153	177	136	110	90
	1000	146	176	243	208	111	91

response closely agree with the formant frequencies of the stimulus. We cannot however, assume that the subject «perceived» the same formant frequencies even if he gives the same response vowel to two stimuli which differ in pitch. Therefore, the formant frequencies of the average responses should not be compared between stimuli differing in pitch, and our material cannot elucidate the relationships between the stimulus spectra differing in pitch and the «perceived» formant frequencies.

A comparison of the mean formant frequencies of the response vowels obtained with and without vibrato would inform us on the effect that the vibrato may have on the extraction of sensory characteristics, as mentioned. Such comparisons can be made in Fig. 2. It shows the difference in the mean formant frequencies of the response vowels. Negative values indicate that the mean dropped when the stimulus had vibrato. Differences exceeding two times the standard error of difference were found in 13 cases (circled dots in Fig. 2). These cases pertain to 10 different pairs of stimuli. Thus, under certain

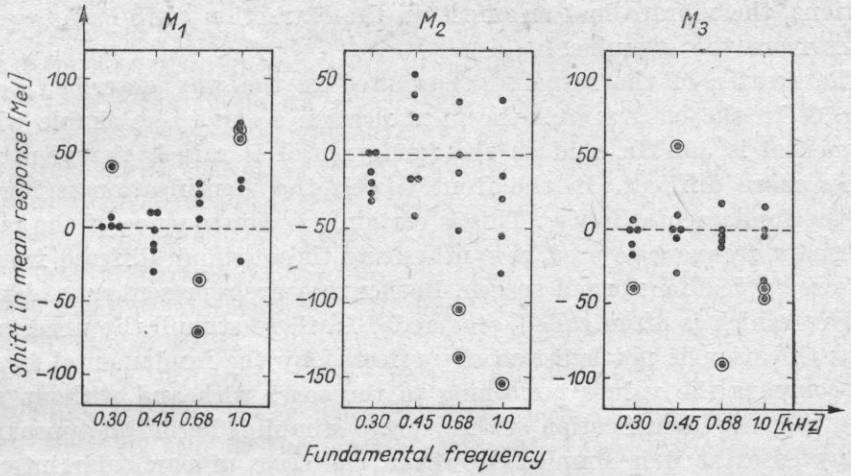

Fig. 2. Difference in the first (left), second (middle), and third (right) formant frequencies in Mels of the response vowels collected from a given stimulus when it was presented with and without vibrato. The circled dots represent differences which exceed twice the standard error of difference

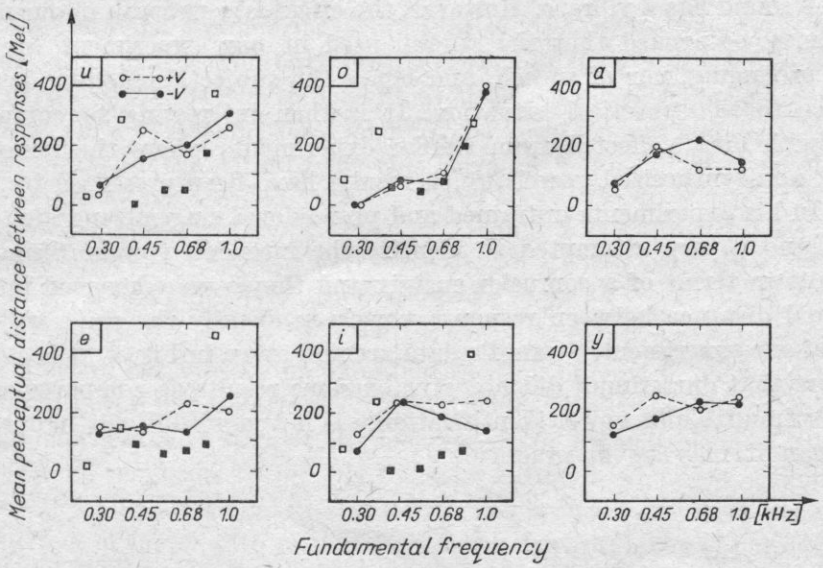

Fig. 3. Scatter of response vowels in terms of the average distance between the individual responses and the average response. Filled and open circles pertain to responses collected when the stimulus was presented with and without vibrato, respectively. The squares give values reported by STUMPF (1926). They were observed when subjects identified vowels sung by two untrained singers (open squares) and one professional soprano (filled squares)

conditions, the vibrato had an effect on the extraction of the sensory characteristics from the stimulus.

The scatter of the responses estimated in the way described is shown in Fig. 3. In the back vowels the identification seems to be simple when the fundamental is 300 Hz, and as the fundamental is raised, the identification becomes more difficult. In the front vowels the identification is hard even when the fundamental is low. This is certainly a consequence of the fact that the formant frequencies used in synthesizing these stimuli differed from those which are typical for normal speech. Rather they are representative of the type of singing which is often called «covered». In these stimuli the uncertainty of the identification is not substantially affected by the fundamental frequency. If we compare the values pertaining to the cases with and without vibrato, we find that the identification of the vibrato stimuli is harder in approximately as many cases as it is simpler. However, the cases of particular interest are those where the responses changed significantly in one or more formant frequencies when the vibrato was added to the stimulus. This happened in 10 cases, and in these cases the sensory characteristics can be assumed to have changed owing to the vibrato. Out of these 10 cases the identification became more uncertain in 7 cases. This seems to support the assumption that the vibrato may have a slight effect on the difficulty with which a sound is identified as a vowel, and the effect is that the identification is somewhat harder to perform when the sound has a vibrato. However, the effect is very weak in our material.

It may be argued that the stimuli used in our experiment are typical neither of singing, nor of speech, and hence the subjects' reactions have little relevance to the practical situation. It is then interesting to compare our results with data collected from similar experiments where the stimuli were natural, non-synthesized vowels. Such results have been presented by STUMPF (1926). In his experiments untrained and professional singers sang high-pitched vowels, and a jury attempted to identify the intended vowels. Stumpf gave his results in terms of a confusion matrix and they were converted into mean perceptual distance between response vowels in exactly the same way as the results of our experiment. Stumpf's data are also given in Fig. 3. It is clear from the figure that our stimuli did not give extreme results as compared with the natural stimuli. Thus, vowel identification was not more difficult in our experiment than it may be in practice.

5. Discussion

Our experiments have shown that the vibrato may have a slight effect on vowel identification. It occasionally changes the interpretation of the acoustic stimulus somewhat and it seems to obscure the vowel identity slightly. This result may seem rather unexpected. However, it is in agreement with previous findings of CARLSON et al. (1973). They studied the vowel identification at the

boundary between [e] and [i], and used a constant as well as a gliding fundamental frequency. The vowel identification seemed to be somewhat harder to perform when the fundamental was gliding. This appears to suggest that the difference limen for formant frequency is somewhat larger when the fundamental is not kept constant (cf. FLANAGAN 1955). The primary question seems to be to what extent we can derive information on the formant frequencies of a vowel sound from the phase relation between the frequency and the amplitude of the spectrum partials; if the amplitude of a partial rises with increasing pitch, this means that this partial is lower in frequency than the formant, and vice versa.

The effect of the vibrato was found to be very slight in our experiment. It is probably related to the extraction of sensory characteristics of the stimulus. In some stimuli the determination of what is characteristic may be very simple regardless of whether or not the stimulus has a vibrato. In other stimuli it may be hard to determine what is characteristic and in these cases the vibrato is more likely to play a role. Probably a singer would be intuitively aware of this possibility and make systematical use of it in singing. Then, the effect of the vibrato might be greater in practice than in our experiment.

According to our results, the vowel identification is somewhat harder to perform when the stimulus has a vibrato. This is to say that the vibrato tends to conceal the singer's formant frequencies. But why should a singer try to conceal the formant frequencies? Substantial deviations from the formant frequencies of normal speech have been observed at professional singers (SUNDBERG 1970; 1975). It may be important to obscure the consequences for vowel quality that such deviations may lead to, and professional singers may make systematical use of the vibrato for this purpose.

6. Conclusions

The effect of the vibrato on the identification of vowels seems to be very slight and to occur under certain conditions only. However, the vibrato may change the vowel quality slightly. In such cases, the identification of the sound as a specific vowel seems to be slightly more difficult when the stimulus has a vibrato. It is believed that a singer may in practice profit systematically from this effect of the vibrato so as to reduce the perceptability of her deviations from the formant frequencies of normal speech.

Acknowledgement. Stefan PAULI is acknowledged for his assistance in computer programming. The work was supported by The Bank of Sweden Tercentenary Foundation.

References

[1] R. Carlson, G. Fant, B. Granström, *Two-formant models, pitch and vowel perception*, paper presented at the Symposium on Auditory Analysis and Perception of Speech (Leningrad 1973), Academic Press, London 1975.

[2] G. Fant, *Speech sounds and features*, MIT-Press, Cambridge, Mass., 1973, p. 36 and p. 84.

[3] J. L. Flanagan, *A difference limen for vowel formant frequency*, J. Acoust. Soc. Am., **27**, 3, 613 (1955).

[4] B. Lindblom, *Experiments in sound structure*, paper presented at the 8th Int. Congr. of Phonetic Sciences, Leeds, Aug. (1975).

[5] R. Plomp, *Timbre as a multidimensional attribute of complex tones*, in *Frequency Analysis and Periodicity Detection in Hearing*, ed. by R. Plomp and G. F. Smoorenburg, A. W. Sijthoff, Leiden 1970, p. 397.

[6] A. W. Slawson, *Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency*, J. Acoust. Soc. Am., **43**, 1, 87 (1968).

[7] C. Stumpf, *Die Sprachlaute*, Springer, Berlin 1926, p. 77–85.

[8] J. Sundberg, *Formant structure and articulation in spoken and sung vowels*, Fol. Phon., **22**, 1, 28 (1970).

[9] J. Sundberg, *Pitch of synthetic sung vowels*, STL-QPSR, 1, 34 (1972).

[10] J. Sundberg, *Formant technique in a professional female singer*, Acustica **32**, 3, 90 (1975).

Received on 5th July 1976
Revised 12 May 1977