AUDITORY AND ACOUSTIC ANALYSIS OF VOICE QUALITY VARIATIONS IN NORMAL VOICES

Elisabeth Zetterholm
Dept. of Linguistics and Phonetics, Lund University, Sweden

ABSTRACT

Anatomical differences and habitual long-term features are components in the individual voice quality and are cues about the speakers identity for the listener. Terms for voice quality or phonation types for use in normal speech often come from studies of pathological speech (laryngeal setting). Therefore they may not be well suited to describe voice quality (especially the variations) of a normal voice. A part of a case study of impersonation, focusing on the imitation of the voice and the speech behaviour, concentrating on voice quality, is presented. This paper also shows some reference types of voice qualities, recorded by a trained phonetician, and their acoustic correlates. An analysis of normal voices are also presented and the aim of that study is to describe normal voices and to use appropriate perceptual voice quality parameters and to find acoustical correlates to the auditory impression.

1. INTRODUCTION

The voice quality, depending on anatomical differences and a person's phonetic habits, are parts of our identity. In *The encyclopedia of language and linguistics* [5] the author says that "voice quality refers to those features of speech which are present more or less all the time that a person is speaking". The features are cues about the speaker's identity, sex, age, social aspects and regional origin, for the listener. The anatomical component in the voice quality, such as the size of the vocal tract, is outside of our control, but there is also a component in the speaker's voice quality that is learned.

We use different voice qualities in normal speech depending on the situations, but there are also linguistic distinctions in some languages. Creaky voice is one common quality with linguistic distinctions, as in some languages in Africa where it is used for phonological contrast to distinguish types of consonants and vowels, from sounds with normally voiced phonation. Nasal voice also has a linguistic distinction, in different languages, Hindi and Yoruba for example [11, 15].

Prosodically voice quality can be used as a boundary signal, and for example creaky voice is an "end-of-utterance" phenomenon [1, 7, 15]. Variation in voice quality is then typically used together with other phonetic cues as final lengthening, drop in F0, decreased intensity and possibly pausing.

Wardhaugh [23] talks about "human speech patterns" and discusses if there are some unique sets of items or patterns, which make it possible to identify regional and social dialects. Elert and Hammarberg [3] have researched some Swedish dialects and found out that there are obvious differences between the regional dialects in both average F0 and voice quality. Interesting studies have been done of social markers in speech focusing both on grammatical features, prosody and voice quality [4, 7, 10, 14, 16, 19, 20, 21]. Results have shown that phonological systems, pronunciations and how often they are used differ between social groups and social status often correlates with a difference in laryngeal settings. Creaky voice is often said to be a male characteristic voice quality and breathy voice seems to be more common among female speakers [17]. Nasal voice is sometimes correlated to high social prestige, in RP English for example.

Voice quality is used paralinguistically in attitudes and emotions [12, 18, 25]. To signal bored resignation a creaky voice is used in the RP dialect in English and breathy voice appears to be signalling intimacy in many languages [13, 15, 17]. In many cultures harsh voice is used paralinguistically as a signal of anger and aggression [13, 15, 17].

2. IMPERSONATION: A PHONETIC CASE STUDY OF THE IMITATION OF A VOICE

This section of the paper presents a phonetic case study of impersonation, focusing on the imitation of the voice and the speech behaviour. Only *one* impersonator and how he works with *one* of his impersonations has been studied. The study is restricted to phonetic aspects and ignores other aspects such as non-verbal, extra-linguistic features, e.g. body language. The aim of the study is to ascertain what particular aspects of the impersonator's voice and speech behaviour are being manipulated during an act of impersonation.

In the present paper I will concentrate on aspects of the impersonation study related to voice quality. For further information see [24].

For this analysis, the Swedish impersonator Göran Gabrielsson made recordings of speeches by three well-known persons. Recordings of the original voices were taken from public appearances, and Gabrielsson recorded the same speech material twice, with his own voice, and as an impersonation. These recordings were made in Gabrielsson's own studio.

In this paper the analysis is limited to the recording of Carl Bildt, a well-known politician. This impersonation was the best one and requires a very critical listening to detect the impersonation, depending on speech style, the voice and speech behaviour changes. Two recordings of Carl Bildt were used for the analysis, an interview and a political speech. The texts were chosen by Gabrielsson. Each recording lasts about 30 seconds.

Usually, when the impersonator works with his impersonations he does not always use exactly the same words as the target speaker. His imitation resembles a caricature of the target speaker's voice and speech behaviour. But for this analysis, to compare the three recordings, it was important to use exactly the same words and not just the same topic as the target speaker.

These recordings have been analysed using both an auditory and an acoustic analysis. In the auditory analysis I have described the phonetic differences between the three recordings. Using the SoundScope program I made an acoustic analysis in which I studied selected aspects of F0, duration and sound spectrum in

the three different recordings and compared them to each other. Those things which I found to be distinctive in the auditory analysis were followed up in the acoustic analysis.

There are obvious differences in the two speakers' voices and speech. Gabrielsson and Bildt have different dialects. The original (Bildt) uses a dialect from the southwest of Sweden, but he now lives in Stockholm and his dialect is affected by the dialect of the latter area. The impersonator lives in Norrköping and represents a more central Swedish dialect. He is a professional impersonator.

2.1. Voice quality

To succeed with the impersonation it seems to be important to imitate the voice quality. In the auditory analysis Bildt's voice seems to be thinner and not as sonorous as Gabrielsson's own voice. The average pitch level is higher for Bildt except at the end of a phrase, when his voice is creaky.

There are some differences in the voice quality between the two recordings of Bildt. In the political speech his voice sounds tenser and louder than in the interview. The same differences can be found in Gabrielsson's impersonation, but not in the recordings with Gabrielsson's own voice.

Gabrielsson is very close to the voice quality of Bildt. Sometimes it is possible to hear the impersonator's own voice in the interview, but in the political speech it sounds more similar to the target speaker. Of course the impersonator tries to trick the listeners when he exaggerates some characteristic features, but he really does change his articulatory settings, the supralaryngeal settings as well as the phonatory settings [8, 13]. The imitation of the prosody is so good that the undershoots in the voice quality seem unimportant for the audience.

In the auditory analysis it is clear that Gabrielsson changes his own voice quality in order to imitate Bildt. I expected to find some differences in the acoustic analysis, in the spectral slope for instance, but the differences were insignificant, as was also the case for jitter and shimmer.

2.2. Vowels

To find out if there were any important differences in the stressed vowels the formant frequencies, F1, F2 and F3 have been measured in the spectra in the recordings of the interview. Generally the formant frequencies are higher in the recordings of Gabrielsson's own voice. There are 39 stressed vowels and 21 of these have a higher formant frequency and 6 have a lower F3 in the recordings with Gabrielsson's own voice than in the impersonation. F2 is higher for 22 of the vowels and lower for 5 vowels in the recording of Gabrielsson's own voice. The duration of the stressed vowels are longer for 19 vowels and shorter for 8 of the 39 vowels in the recording of the impersonation compared with Gabrielsson's own voice. The acoustic values of the impersonation are closer to the values of Bildt's voice than to the impersonator's own voice.

2.3. Concluding remarks of the case study of impersonation

In these analysed recordings the impersonator succeeds with his impersonations, with the prosody as well as with the voice quality. There are some differences between the interview and the political speech. Sometimes it is possible to hear Gabrielsson's own voice in the impersonation of the interview. It is possible that it is easier to exaggerate an imitation with more emphasis and a loud voice, like the political speech.

It is hard to describe voice quality, especially the variations of a modal voice, so it is hard to know exactly what the impersonator really does to change his voice. Maybe he raises the larynx to get the timbre and the raised pitch. To lower the formant frequencies of the vowels, labial protrusion and lip-rounding, or lowering of the larynx, is one possible means [17].

3. TYPES OF VOICE QUALITY

Terms for voice quality or phonation types for use in normal speech often come from studies of pathological speech (laryngeal settings) and it is hard to describe voice quality, especially the variations within the range of a normal voice. Articulatory settings and voice quality settings and their acoustic correlates have been studied by, among others, Catford [2], Honikman [8], Laver [13, 14, 15] and Hammarberg [6]. Hoarseness in children has been studied [22] and de Krom [9] has studied the acoustic correlates of breathiness and roughness. One of his concluding remarks is that "more knowledge about the nature and magnitude of voice quality variations as they occur in the conversational speech of individual speakers" is important for voice quality evaluation. In the description of voice quality it is important to consider both laryngeal and supralaryngeal settings.

The most common types of phonation types (laryngeal settings) by Laver [13] are:

- modal or normal voice
- breathy voice, as opposed to tense or strained voice, with a high rate of air-flow
- creaky voice with very low frequency and usually irregularly spaced in time
- harsh voice with a normal fundamental frequency but aperiodicity or noise in spectrum
- tense or strained voice with a low rate of air-flow (often described like a 'metallic voice')

There are also a number of compound phonation types and the combined breathy and tense voice is an example in my reference types. In spite of the suitability of these terms we still lack a more complete typology for description of voice quality in normal speech.

Recordings of different voice qualities (phonation types) have been made by a trained Swedish phonetician. There are obvious differences already in the waveforms for the six voice qualities (modal, breathy, creaky, harsh, tense and the compound breathy and tense voice). These represent reference types for my study and are shown in figures 1-6 (the same utterance).

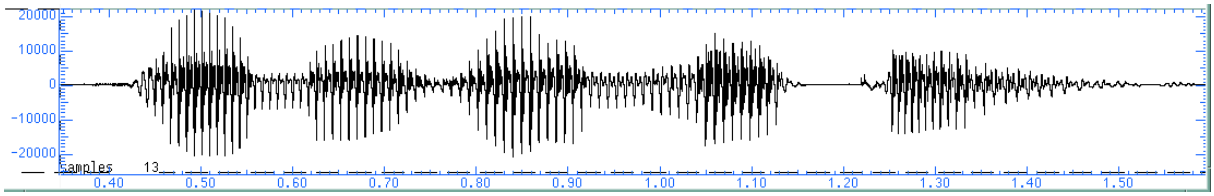

Figure 1: Modal voice

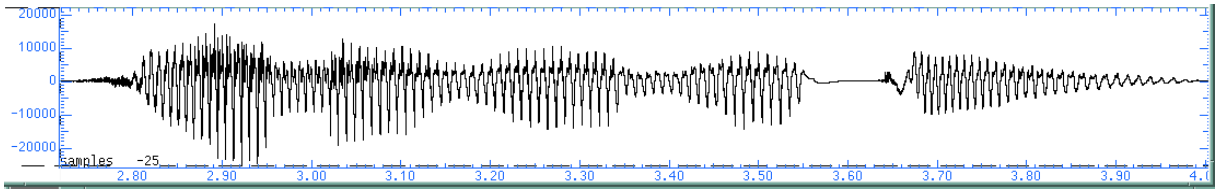

Figure 2: Breathy voice

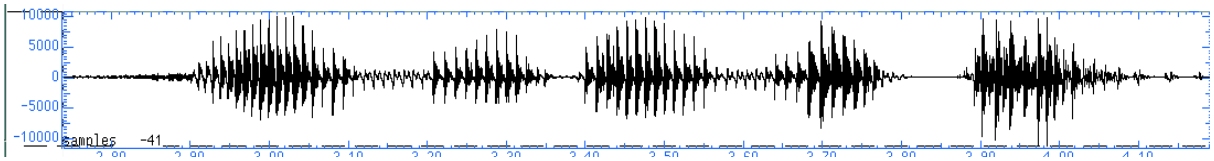

Figure 3: Creaky voice

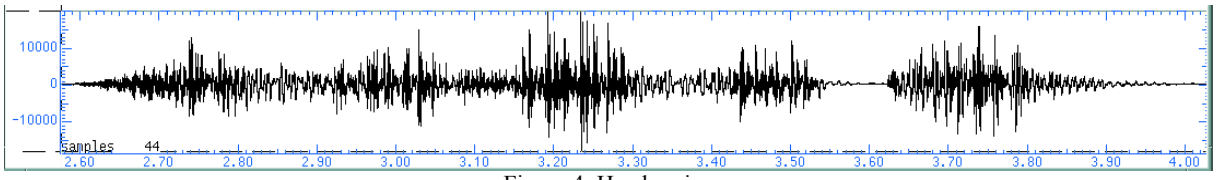

Figure 4: Harsh voice

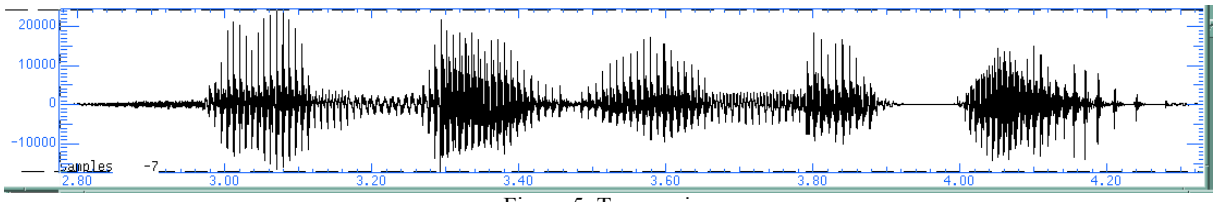

Figure 5: Tense voice

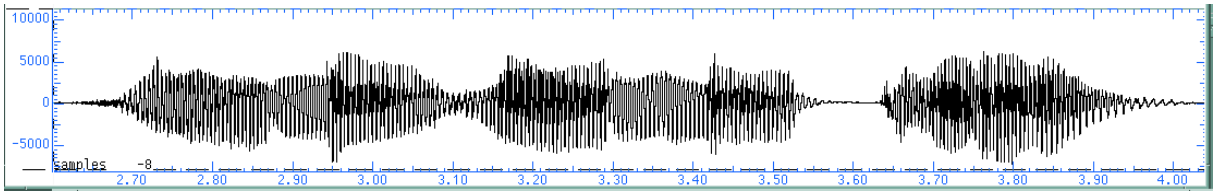

Figure 6: Compound breathy and tense voice

4. ANALYSIS OF NORMAL VOICES

4.1. The aim of the study

In this study I have focused on the voice quality variations of normal voices, the individual voice. The aim of this study is to describe the voices and to use appropriate perceptual voice quality parameters for normal voices and to find acoustical correlates to the auditory impression.

This is a part of a larger study and the results are preliminary.

4.2. Recordings

For the recordings and the analysis, people with normal voices, both men and women, i.e. with no known pathology or special training, have been asked to read some short texts. They have been recorded with a DAT-recorder in a studio at the department of Linguistics and Phonetics in Lund.

4.3. Informants

Twelve women and eight men, age about 25-50 year, were recorded. For the present study I decided to use six voices, three men and three women, for a detailed analysis and another six voices for comparison. The voices I have chosen represents different types of voice quality in the auditory analysis in view of the pitch and a general impression of "soft", "tight" or "thin" voice quality. The similar voices, in the auditory analysis, I put into different categories and compared them to each other, and to the voices in the other categories. They are not matched in relation to age or dialect. My intention was to find some similar acoustic features in these voices. They all live in the south of Sweden but they have different dialects, which are all more or less influenced by the dialect of Lund.

4.4. Analysis

The recordings have been analysed both auditorily and acoustically. In the auditory analysis the voices are described with terms used for classification of pathological voices and also in a way using labels for describing the analyst's impression of the different voices. The auditory impressions are compared to the acoustic analysis in fundamental frequencies, waveforms and spectrograms, using the analysis programme ESPS/Waves+. The waveforms are also compared with waveforms of my reference voices.

The voices in these recordings represent a great variety of normal voices. In describing the differences in normal voices a useful voice quality terminology is required. The parameters for pathological voices are not applicable even if some features, such as creaky, breathy and nasal voice qualities, occur. The adjectives used for a perceptual description are not equivalent and have no consistent acoustic-phonetic description.

In this analysis the auditory impression of pitch level corresponds to the acoustic analysis and the terminology consisting of adjectives seems to correlate as well, for example "thin", "soft", "lax" and "muffle" voice quality, in relation to pitch level. In the waveform it is possible to see some features of voice quality, such as creaky voice. Even if long-term settings of voice quality are meant to be characteristic, individual timing differences may also be important for the impression of the voice quality. In the spectrograms of, for example, the vowel [a] that have been analysed there are differences depending on the dialect, but also differences in formant frequencies and levels, periodicity and noise that may be acoustical correlates to the auditory impression.

It is a challenge to find acoustical correlates to the auditory impression, at least in these recordings, and the results of these analyses are too preliminary for an acoustic description of voice quality in normal speech. Perceptual estimation has to be done and the perceptual parameters in this analysis have to be tested on a group of listeners.

REFERENCES

[1] Bruce, G. 1998. *Allmän och svensk prosodi.* Praktisk lingvistik 16, Department of Linguistics and Phonetics, Lund University.

[2] Catford, J.C. 1964. Phonation Types: The Classification of Some Laryngeal Components of Speech Production. In Abercrombie, D. et al (eds), *In Honour of Daniel Jones*. Longmans, London.

[3] Elert, C. and Hammarberg, B. 1991. Regional Voice Quality Variation in Swedish. In *Proc XIIth Intern Congr Phonetic Science*, Aix-ex-Provence. Frankrike.

[4] Esling, J. 1978. *Voice Quality in Edinburgh: A Sociolinguistic and Phonetic Study*. Unpublished doctoral dissertation. Edinburgh University, Edinburgh.

[5] Esling, J.H. 1994. Voice Quality. In Asher, R.E. and Simpson, J. M.Y. (eds), *The encyclopedia of language and linguistics.* Pergamon Press, England.

[6] Hammarberg, B. 1986. *Perceptual and Acoustic Analysis of Dysphonia*. Studies in Logopedics and Phoniatrics No. 1. Huddinge University Hospital, Stockholm.

[7] Henton, C. and Bladon, A. 1988. Creak as a Sociophonetic marker. In Hyman, L. and Li, C. (eds), *Language, speech and mind. Studies in honour of Victoria Al. Fromkin*. Routledge, London.

[8] Honikman, B. 1964. Articulatory settings. In Abercrombie, D. et al. (eds), *In Honour of Daniel Jones*. Longmans, London.

[9] de Krom, G. 1994. *Acoustic Correlates of Breathiness and Roughness. Experiments on voice quality*. Research Institute for language and Speech. Utrecht University.

[10] Labov, W. 1974. Logiken i "Nonstandard English". In Loman, B., *Barnspråk i klassamhälle*. LiberLäromedel, Lund.

[11] Ladefoged, P. and Maddieson, I. 1997. *The Sounds of the World's Languages.* Blackwell Publishers Ltd, Oxford.

[12] Laukkanen, A-M., Vilkman, E., Alku, P., Oksanen, H. 1997. On the perception of emotions in speech: the role of voice quality. *Journal of Logopedics and Phoniatrics Vocology,* 22 (4), 157-168.

[13] Laver, J. 1980. *The Phonetic Description of Voice Quality.* Cambridge University Press.

[14] Laver, J. 1991. *The Gift of Speech. Readings in the Analysis of Speech and Voice*. Edingburgh University Press.

[15] Laver, J. 1994. *Principles of phonetics*. Cambridge University Press.

[16] Laver, J. and Trudgill, P. 1979. Phonetic and linguistic markers in speech. In Scherer, K.R. and Giles H. (eds), *Social markers in speech*. Cambridge University Press.

[17] Lindblad, P. 1992. *Rösten*. Lund: Studentlitteratur.

[18] Mozziconacci, S. 1998. *Speech Variability and Emotion: Production and Perception*. Technische Universiteit Eindhoven.

[19] Pittam, J. 1994. *Voice in Social Interaction.* SAGE Publications, California.

[20] Pittam, J. and Scherer, K.R. 1993. Vocal Expression and Communication of Emotion. In Lewis and Haviland, J.M. (eds), *Handbook of emotions*. The Guilford Press, New York.

[21] Scherer, K.R. 1979. Personality markers in speech. In Scherer, K.R. and Giles, H. (eds), *Social markers in speech*. Cambridge University Press.

[22] Sederholm, E. 1996. *Hoarseness in ten-year-old children. Perceptual Characteristics, Prevalence and Etiology.* Studies in Logopedics and Phoniatrics No. 6. Huddinge University Hospital, Stockholm.

[23] Wardhaugh, R. 1998. *An Introduction to sociolinguistics.* Blackwell Publishers, Oxford.

[24] Zetterholm, E. 1997. Impersonation: A Phonetic Case Study of the Imitation of a Voice. In Svantesson, J.-E. (ed), *Working Papers 46*. Lund University.

[25] Zetterholm, E. 1998. Prosody and voice quality in the expression of emotions. In Mannell, R. and Robert-Ribes, J. (eds), *Proceedings of the Seventh Australian International Conference on Speech Science and Technology*. Sydney.