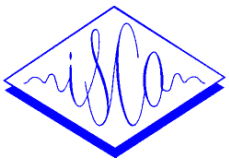

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PERCEPTION OF APERIODIC SPEECH SIGNALS

Dieter Huber

Department of Information Theory
Chalmers University of Technology
S-412 96 Gothenburg
Sweden

ABSTRACT

This paper addresses the perceptual relevance of laryngealization as a potential boundary cue in continuous speech utterances. It investigates, in other words, the problem whether the short-time occurrences of various patterns of aperiodic voice vibration frequently found at boundary locations in human speech can actually be perceived and discriminated by human listeners in normal communicative situations, and thus may be taken to contribute (1) to the signal information needed for the correct recognition and interpretation of the structural properties of the message, and (2) to the impression of naturalness in human versus computer speech. Four patterns of laryngealization have been examined systematically: *glottalization, creaky voice, creak* and *diplophonic phonation*. The results of this study indicate that human listeners evidently exploit aperiodicity in the acoustical speech signal for segmentation but not for classification purposes.

1. INTRODUCTION

The basic characteristics of pitch perception in speech communication have over the years been extensively studied within the fields of psychoacoustics, phonetics, auditory physiology, and various realms of speech science and technology (see [2],[6],[9] and [10] for overviews). Consequently, it is today well established that the way in which human listeners perceive F_0 values and F_0 fluctuations in the acoustical speech signal depends on a number of strongly interrelated factors, including for instance the F_0 average, the range and direction of F_0 movements, signal duration and intensity, the rhythmical properties of the utterance, the spectral composition of the transmitted signal, and various potentially distortive aspects such as masking, binaural listening, and fatigue.

Most of these findings are based on the study of pitch perception in stationary, quasi-periodic speech signals. Non-stationary, aperiodic signals, i.e. exhibiting occasional time and/or amplitude irregularities between successive periods of glottal excitation, have traditionally been disregarded and considered as reflecting pathological voice phenomena which are of no immediate concern to normal speech processing applications. However, not only does aperiodic voice vibration occur far too often in normal, nonpathological voices of both women and men to be classified simply as a clinical syndrome or voice disorder (e.g. [6] and [9]). Aperiodic phonation has also been shown to be systematically employed by human speakers as an important demarcation cue in connected speech, used to support the segmentation of the continuous speech utterance into communicatively relevant information units (e.g. [5],[6] and [11]). The question remains, however, to what extent human listeners are able to perceive these cues and to use them in order to decode the intended meaning structure of the transmitted message.

2. LARYNGEALIZATION

In an earlier study of voice phenomena in Swedish intonation based on three texts (one narrative, one descriptive and one argumentative, comprising a total of 2610 running words, 176 sentences and 65 paragraphs), read by four native speakers of Swedish (two female/two male; two radio journalists/two experienced public speakers) and recorded in an anechoic, sound-insulated environment using high-quality digital recording equipment (SONY PCM-F1), four different patterns of aperiodic glottal excitation have been observed to occur consistently at different kinds of text junctures [6]. These patterns, commonly referred to as *laryngealization* in the research literature on speech production and perception, are exemplified below in figure 1 and will be denoted in the following as *glottalization, creak, creaky voice* and *diplophonic phonation*.

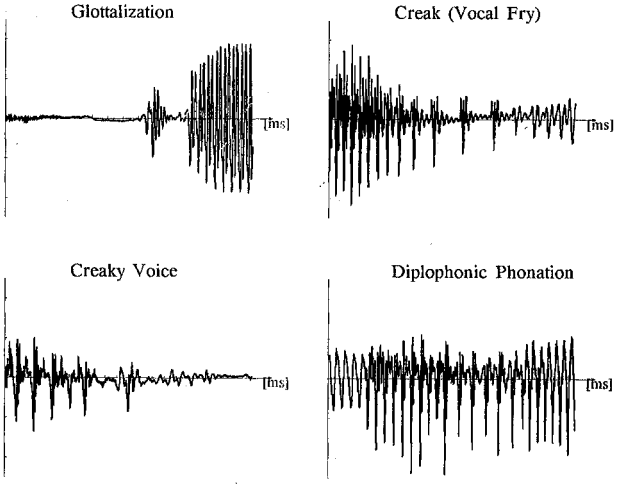

Figure 1 Patterns of Aperiodic Voice Vibration

Glottalization as defined in this study is characterized in the acoustical domain by one initial vibratory cycle that is clearly demarcated from the ensuing regular periodic vibrations by a markedly lower amplitude and a significantly longer vibration period. Most typically in our material, the time period for the first *glottalized* cycle is approximately twice (for the male speakers) or three times (for the female speakers) longer than the vibration periods in the ensuing cycles of quasi-periodic phonation. The term glottalization for this kind of phenomenon was chosen for its connotations with the terms *glottal* (thus defining it as a phonatory rather than as an articulatory

10.21437/ICSLP.1992-162

gesture) and *glottal stop* (with which *glottalization* shares common grounds in the production domain).

The term **creak** refers to a pattern of glottal vibration consisting of a train of discrete low-frequency pulses at a repetition rate (frequency) typically in the range between 20Hz and 50Hz, which is produced with nearly complete damping of the vocal tract between successive excitations. The typical vibratory pattern in *creak* phonation thus displays a rapidly opening and closing motion of the vocal folds and a very long closed period, with individual glottal cycles so long and so highly damped that individual pulses are perceived rather than a fused tonal quality. Alternative names often used in the literature for this pattern of glottal excitation include vocal fry, pulse register and strohbass.

Creaky Voice patterns found in our material are characterized in the acoustical domain by clearly visible period-to-period variations in fundamental frequency that are often accompanied by more or less pronounced period-to-period fluctuations in intensity (amplitude). Contrary to the small-scale cycle-to-cycle variations normally occurring during quasi-periodic phonation in sustained speech, which are commonly referred to in the literature as *jitter* and *shimmer*, aperiodicity in creaky voice is considerably more pronounced, viz. both frequency as well as amplitude changes from one cycle to the next are more extreme. In addition to aperiodicity and fluctuations in intensity, even a component of *breathiness* and/or *partial devoicing* can sometimes be observed in creaky voice phonation (see also [12] and [13]). Contrary to full creak, the aperiodic cycles observed in creaky voice phonation are typically composed of individual glottal pulses that look superficially similar to those normally found in modal voice.

With the term **diplophonic phonation** we refer to a pattern of voice vibration that is characterized by an alternation between strong and weak glottal excitations during phonation, which is typically sustained over longer stretches of speech. A subcategory of diplophonic phonation, sometimes referred to in the literature as *dicrotic dysphonia*, consists of a vibratory pattern in which the vocal cords separate twice in quick succession and then approximate firmly in a relatively long closed phase. It must be noted in this context that our use of the terms *diplophonic phonation* and *diplophonia* is not to be confused with the use of these terms sometimes found in papers on phoniatric research, where they refer to the rare ability of some speakers to phonate at two distinct levels of periodicity at the same time.

Each of these patterns of aperiodic voice excitation has been systematically investigated and described in earlier publications with respect to (i) their acoustical characteristics and (ii) their distributional properties between different speakers (female and male), different languages (Swedish, English and Japanese), different speaking styles, and different kinds of text boundaries (cf. [3],[5],[6],[7] and [14]). Similar results for other languages have also been reported in references [8] for German and [15] for Japanese and English dialogues. In reference [3] a new pitch determination algorithm was presented which outperforms standard PDA methods such as for instance the Gold-Rabiner algorithm, the SIFT algorithm and the cepstrum method with respect to correct voicing decisions and pitch estimations in both quasi-periodic and aperiodic speech signals.

The present paper continues this line of research on the communicative function of aperiodicity in human speech, and addresses the question of the perceptual relevance of laryngealization as a potential boundary cue in continuous speech utterances. It investigates, in other words, the problem whether these four patterns of aperiodic voice vibration can actually be perceived and discriminated by human listeners in normal communicative situations, and thus may be taken to contribute (1) to the signal information needed for the correct recognition and interpretation of the structural properties of the message, and (2) to the impression of naturalness in human versus computer speech. Three subproblems are examined systematically:

1 - Can human listeners **perceive** these patterns **at all** in verbal surroundings where they occur only sporadically and normally only for very short time-intervals in otherwise stationary, quasi-periodic speech?

2 - Can human listeners **distinguish categorically** between these four patterns of aperiodicity, i.e. *creaky voice, creak, glottalization* and *diplophonic phonation*?

3 - Can human listeners correctly **recognize** F_0 **movements**, i.e. variations of fundamental voice frequency, during periods of laryngealization, or are they classified as simple either-or-phenomena?

In a larger perspective, this study thus addresses the question to what extent these four patterns of aperiodic voice vibration in the speech signal can actually be used as reliable cues for the **classification** of communicatively significant boundary locations in continuous speech utterances?

3. DATA

The material used for this study was selected from the CTH Speech Database [4] and consists of short (approximately one minute) extracts from each of the three texts (i.e. one narrative, one descriptive and one argumentative) read by the same four native speakers of Swedish (i.e. two female/two male; two radio journalists/two experienced public speakers) used in the earlier studies on the acoustics, distributions and pitch determination of aperiodic speech signals reported in references [3],[6] and [7]. Thus, approximately twelve minutes of recorded speech were examined, containing a total of 416 occurrences of laryngealization at various boundary locations. The distribution of these 416 occurrences between the four different patterns of aperiodic voice vibration is listed below in table I.

Table I Patterns of Aperiodic Voice Vibration in the Test Material.

	Glottalization	Diplophonia	Creaky-Voice	Creak	Total
n	81	109	122	104	416
%	19.4	26.2	29.3	25.1	100.0

It must be appreciated in this context that the speakers were chosen in an attempt to secure a high standard of professionality in oral reading skills and to minimize, as far as possible, potential dialectal and/or sociolectal differences. None of the speakers reported any history of speech or hearing disorders. Further details concerning the speakers, the material, the recording procedures and the signal processing can be found in references [4] and [6].

4. TEST PROCEDURES

For the listening tests, a total of 32 listeners was engaged, including 22 speech scientists and speech therapists (G1) as well as 10 laymen in the field of language research (G2).

Two sets of listening test materials were designed in order to establish both the absolute threshold(s) of aperiodicity **detection** (viz. subproblem 1) and the difference threshold(s) of aperiodicity **discrimination** (viz. subproblems 2 and 3).

For the investigation of the subproblems 1 and 3, variable signal stimuli were included in the texts in order to permit stepwise adaptation in the duration domain (cf. [16]). Variations were controlled by framewise clipping and/or pasting of the natural speech signal, using the 16ms signal analysis window for calibration.

For the investigation of subproblem 2, a fixed set of constant stimuli was presented to the listeners twice in randomized order. A total of 80 such stimuli was prepared, comprising (i) 20 short extracts from the read texts (5 for each speaker) containing one of the four patterns of laryngealization in the natural recordings that were reliably detected by the majority of our subjects during the first test (subproblem 1), plus (ii) three manipulated versions of each of these recordings, where the natural pattern had been replaced with, in turn, the other three patterns by employing a simple cut-and-paste procedure.

Each of the three subproblems was treated as a separate test, run on different days. All tests were conducted in a *forced choice* situation.

5. RESULTS

5.1 Subproblem 1: Detection

Table II below lists the responses of the 32 subjects to the first test, which was to establish whether human listeners can actually **perceive** aperiodic voice vibration **at all** in verbal surroundings where they occur only sporadically and normally only for very short time-intervals in otherwise stationary, quasi-periodic speech? The listeners task was to mark perceived occurrences of laryngealization in the transcribed texts as the were listening to the recordings. No discrimination betweeen different patterns of laryngealization was required at this stage. For reasons of presentational clarity, however, the results in the table are listed separately for each of the four patterns and for the two groups of expert (G1) and non-expert (G2) listeners respectively.

Table II Results of the Detection Test

		Glottalization	Diplophonia	Creaky-Voice	Creak	Total
G1(22)	n	1034	1892	1386	1870	6182
	%	58.1	78.9	51.6	81.7	67.5
G2(10)	n	320	741	573	787	2421
	%	39.5	67.9	46.7	76.0	58.2

These figures clearly indicate that our listeners are indeed able to perceive at least three of these four patterns of aperiodic voice vibration, viz. *glottalization, diplophonic phonation* and *creak*, at above chance level, largely regardless of any duration constraints. The detection results for *creaky voice* are somewhat more ambiguous and appear to depend largely on the distance to the overall F_0 average in the surrounding quasi-periodic speech utterance, whereas changes in duration reveal to have had no significant impact on our listeners judgements. Tests to further corroborate these interdependencies are under way and will be reported in a later paper.

5.2 Subproblem 2: Pattern Discrimination

Table III below shows a confusion matrix which summarizes the results of the second test, which was aimed to establish whether human listeners are able to **distinguish categorically** between these four patterns of aperiodicity, i.e. *creaky voice, creak, glottalization* and *diplophonic phonation*? During this test, the listeners were working with the transcribed material in which the **location** of laryngealization was indicated and they task was to identify which of the four **kinds** of pattern they were actually listening to in the recordings.

Table III Confusion Matrix for the Discrimination Test

		Glott	Dipl	Creaky	Creak
Glott	n	784	208	288	-
	%	61.2	16.3	22.5	-
Dipl	n	144	561	336	239
	%	11.3	43.7	26.2	18.8
Creaky	n	79	368	416	432
	%	6.2	28.6	32.5	33.7
Creak	n	96	224	448	512
	%	7.5	17.5	35.0	40.0

As indicated by these results, at least our listeners do not appear to be able to correctly identify and distinguish between these different patterns of laryngealization in a consistent and reliable manner. For instance, *creaky voice*, as far as it is detected at all (compare the results in subsection 5.1), is mostly confounded with either creak or diplophonic phonation. Given the distribution of these patterns between different kinds of morphologically, syntactically, textually and/or physiologically induced boundaries (cf. [5] and [6]) it must be concluded that even in the cases of predominantly correct identifications (e.g. for *glottalization*) at least some of the listeners apparently identify these patterns not so much by virtue of their acoustical characteristics but rather in a top-down, expectancy-driven way on the basis of their interpretation of surrounding features of linguistic structure and verbal content.

5.3 Subproblem 3: Movement Discrimination

The purpose of this test was to determine whether human listeners are able to correctly **recognize F_0 movements**, i.e variations of pitch frequency, during periods of laryngealization, or whether are they classified as simple either-or-phenomena? In order to establish our subjects' ability to discern F_0 movements during the relatively short stretches of laryngealized speech at boundary locations, they were asked to mark one of the following 5 alternatives for each presented stimulus: (1) rising F_0; (2) falling F_0; (3) level F_0; (1) fluctuating F_0 (including both fall-rise and rise-fall patterns); (5) indeterminable F_0 movement.

The results of this test reveal that the panel of listeners overwhelmingly judged occurrences of laryngealization within the realm of this study (i.e. presented as short-time boundary signals and not as long-time idiosyncratical voice disorder) as predominantly either-or-phenomena without discriminating any further variations. The instances where listeners marked F_0 movements and/or F_0 directions in the stimuli pertain almost exclusively to occurrences of creaky voice at sentence

final boundaries which were accompanied by a distinct fall in fundamental voice frequency, and were noted by members of the subgroup of speech therapists only.

6. COMMUNICATIVE SIGNIFICANCE

The results of an earlier study in the time-alignment between these four different patterns of laryngealization with different kinds of textual, sentential, and prosodic junctures (paragraph, sentence, clause, constituent, intonation-unit onset and offset, speech inhalation pause, etc.) indicated the following general tendencies:

Glottalization is most often used in our material to mark the onset of sentence internal clause boundaries, irrespective of whether they coincided with a speech inhalation pause and/or the onset of an intonation unit.

Diplophonic Phonation occurs predominantly at utterance internal positions in word junctures between adjacent voiced (mostly vowel) sounds, where it apparently serves as a kind of *vocal (laryngeal) hiatus*. At intonation offset locations, diplophonic phonation often co-occurs with other vocal features (e.g. breathiness) or in the transition before a stretch of creaky voice and/or devoicing.

Creaky Voice occurs at all three textual/sentential positions investigated so far in our experiments, i.e. at transparagraph, intersentence, and intrasentence boundaries, however, with a clear majority of incidences at intersentence locations. Creaky voice displays a clear and consistent tendency to time-align with succeeding speech inhalation pauses and has been established predominantly at intonation offset locations with a low terminal F_0 contour.

Creak, on the other hand, occurs mostly at intersentence locations that coincide with intonation unit offsets that are not marked by a low terminal F_0 contour. Time-alignment with speech inhalation pauses proves to be highly speaker dependent, with a markedly lower correlation rate for our female as compared with the male speakers.

Detailed data about these co-occurrences have been presented earlier in [6]. In summary, it appears that all four patterns of aperiodic voice vibration observed in our data serve as genuine speech demarcation mechanisms which are used by our **speakers** quite systematically to signal different kinds and degrees of "punctuation" in coherent speech.

The question asked at the outset of this present investigation in the perception and communicative function of larygealization was to what extent these four patterns of aperiodic voice vibration in the speech signal can actually be used by our **listeners** as reliable cues to the detection and, ultimately, **classification** of potential boundary locations in continuous speech utterances?

The results of this study show that while the occurrence of laryngealization in the speech signal as such appears to provide a reliable cue to the **detection** of boundary locations in continuous speech, the distinction between the four patterns of aperiodic vibration is considerably more diffuse and does not seem to be exploited in any significant way for the **classification** of the kind of boundary. Thus, the strong correlations between "boundary signal" on one side and "boundary class" on the other, reported earlier in [3] and [6] thus did not influence at least our listener's judgement with respect to a finer sub-classification of juncture cues. Human listeners evidently exploit aperiodicity in the acoustical speech signal for segmentation but not for classification purposes.

As regards the importance of these results and their applicability to practical computer speech applications, it is today widely acknowledged that the accurate representation of the voicing characteristics is of paramount importance for all aspects of speech signal processing (synthesis, coding, transmission, compression, enhancement, etc). In speech coding, for instance, the quality of the vocoded speech deteriorates rapidly as a function of imprecise pitch estimates. In speech synthesis, considerable effort is dedicated today to the development and implementation of glottal models for the generation of natural sounding pitch contours in text-to-speech. Equally important, modern ASR systems make increasingly use of the F_0 information inherent in the pitch contours for the segmentation of the continuous speech utterance into semantically meaningful "chunks" and for the automatic detection of the stressed parts of the message. Finally, the voicing behaviour of individual speakers and their F_0 frequency distributions (FFD) calculated over relatively short samples of speech have been shown to provide reliable cues for speaker identification and verification.

Clearly, given the importance of aperiodic voice vibration both as a boundary cue in connected speech and as a social as well as idiosyncratic marker useful e.g. to increase the naturalness in synthesized speech, a more accurate and reliable way to use aperiodic speech signals in various computer speech applications is called for.

REFERENCES

[1] A Fourcin & E Abberton, "Laryngograph Studies of Vocal-Fold Vibration", *Phonetica* **34**, pp.313-315, 1977.

[2] W Hess, *Pitch Determination of Speech Signals*, Springer Verlag, 1983.

[3] P Hedelin & D Huber, "Pitch Period Determination of Aperiodic Speech Signals", *ICASSP'90*, pp.361-364, Albuquerque, 1990.

[4] P Hedelin & D Huber, "The CTH Speech Database: An Integrated Multilevel Approach", *Speech Communication* Vol.9, No.4, pp.365-374, 1990.

[5] D Huber, "Laryngealization as a Boundary Cue in Read Speech", *Second Swedish Phonetics Conference*, pp.66-67, Lund, 1988.

[6] D Huber, "Aspects of the Communicative Function of Voice in Text Intonation", Ph.D. Thesis, Göteborg/Lund, 1988.

[7] D Huber, "Voice Characteristics of Female Speech and their Representation in Computer Speech Synthesis and Recognition", *EUROSPEECH'89*, Vol.2, pp.477-480, Paris, 1989.

[8] A Kießling, "Optimierung des DPGF-Grundfrequenzverfahrens durch besondere Berücksichtigung irregulärer Signalbereiche", Diplomarbeit, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1990.

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

[10] D Klatt, "Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Speech Perception", *JASA* **53**, pp.8-16, 1973.

[11] J Kreiman, "Perception of Sentence and Paragraph Boundaries in Natural Conversation", *Journal of Phonetics* **10**, pp.163-175, 1982.

[12] J Laver, S Hiller & R Hanson, "Comparative performance of pitch detection algorithms on dysphonic voices", *ICASSP'82*, pp.192-195, Paris, 1982.

[13] T V Sreenivas, "Pitch Estimation of Aperiodic and Noisy Speech Signals", Ph.D. Thesis, Tata Institute of Fundamental Research, Bombay, 1981.

[14] D Huber, "On the Discourse Function of Intonation", *XIIth ICPhS*, Vol.V, pp.190-193, Aix-en-Provence, 1991.

[15] D Huber, "Prosodic Structures in Japanese and English Dialogues", *PANASIATIC LINGUISTICS '92*, Vol.1, pp.60-74, Bangkok, 1992.