gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

MÜSLI: A Classification Scheme For Laryngealizations

A. Batliner, S. Burger,
B. Johne, A. Kießling

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Report 45
Dezember 1994

Dezember 1994

A. Batliner, S. Burger,
B. Johne, A. Kießling

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (089) 2180 - 2916
e-mail: `ue102ac@cd1.lrz-muenchen.de`

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

From Proc. ESCA Workshop on prosody, Lund, 1993, pp. 176–179.

MÜSLI: A Classification Scheme For Laryngealizations

A. Batliner[1], S. Burger[1], B. Johne[1], A. Kießling[2]

[1] L.M.-Universität München, Institut für Deutsche Philologie,
Schellingstr. 3, 80799 München, F.R. of Germany

[2] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, 91058 Erlangen, F.R. of Germany

ABSTRACT

We developed a classification scheme for laryngealizations that can be used to discriminate the many different shapes of laryngealizations with different feature values. Potential applications are phonetic transcription and automatic detection. The scheme was developed and tested with a database from 4 speakers that contains more than 1200 laryngealizations.

INTRODUCTION

The normal speech register (modal voice) comprises a F0-range from about 60 to 250 Hz for male speakers and a F0-range from about 120 to 550 Hz for female speakers. Below this register there is a special phonation type whose mechanisms of production are not totally understood yet and whose linguistic functions are little investigated until now. There is a variety of different terms for this phenomenon, which are used more or less synonymously: creak, vocal fry, creaky voice, pulse register, laryngealization, etc. We use "laryngealization" (henceforth LA) as a cover term for all these phenomena that show up as irregular, voiced stretches of speech. Normally, LAs do not disturb pitch perception but are perceived as suprasegmental irritations modulated onto the pitch curve. Although LAs can be found not only in pathological speech, but also in normal conversational speech, most of the time, they were not objects of investigation, but considered to be an irritating phenomenon that has to be discarded. On the other hand, recently the fact that LAs often occur at word or morpheme boundaries and thus could be used in speech recognition, has been realized. Efforts for their investigation and classification have been undertaken [2] [3]. In the time signal LAs can look quite different (cf. figure 1–6) and it is not far-fetched to claim that the only common denominator of the different types is their irregularity. LAs can be produced with different means and different states of the glottis but it is not clear yet whether there is a regular relationship between different production mechanisms and different types of LAs showing up in the time signal. In [2] four different types of LAs are characterized (cf. below). We will follow another approach and use non-binary features for our description scheme that can be used by different transcribers in a consistent way. An overspecification can be reduced in a second step. It should be possible to extract the features automatically with standard pattern recognition algorithms.

MATERIAL

We investigated a database of 1329 sentences from 4 speakers (3 female, 1 male; 30 minutes of speech in total). One third of the database consists of real spontaneous utterances gained from human-human clarification dialogues, the rest consists of the same utterances read by the same speakers nine months afterwards (own utterances and partners utterances). Recording conditions were comparable to a quiet office environment. The utterances were digitized with 12 Bit and 10 kHz; for more details cf. [1]. Two trained phoneticians classified the voiced passages as [+/− laryngealized] with the help of a segmentation program (time waveform presented on the screen and iterative listening to the segmented part). 4.8% of the speech in total (7.4% of the voiced parts) were laryngealized (henceforth la).

The mean duration of the LAs was 64.1 ms with a standard deviation of 35.1; minimum = 12.8 ms (1 frame), maximum = 332.8 ms (26 frames). 16% of the LAs extend through a phoneme boundary. The non-la passages will not be considered in this paper. The la parts were plotted with their non-la context and a constant resolution, and a group of 6 experts tried to cluster a subset of these plots manually using different criteria (the 4 classes in [2] as well as phoneme-, context-, and speaker- specific peculiarities). Based on the similarities between the tokens within the clusters and the dissimilarities between tokens of different clusters respectively, several features were chosen for characterizing the LAs adequately. Afterwards, a classification scheme was developed heuristically and subsequently tested and verified with the whole material.

THE CLASSIFICATION SCHEME FOR LARYNGEALIZATIONS

In MÜSLI (**Mü***nchner* **S***chema für* **L***aryngalisierungs*–**I***dentifikation*) six different features in four different domains (cf. table 1) are used for describing LAs. The values of these features can be determined independently from each other and are coded with integers within the ranges from 1–3 or from 1–4. Thus, every LA is determined by a sextuple of integers. In this paper, we will deal only with these features and not with other, e.g. speaker- or context-specific phenomena. Due to the lack of space, not every feature value can be illustrated in figure 1–6, but some of the values can be seen in the captions. The features and their values that are given in brackets are described in the following. In parentheses, the percentage of cases of all LAs assigned to the specific value are given. Values that can probably be combined into one single value (i.e. reduction of overspecification), are given in curly brackets at the end of the description of each feature. Reasons for combining are: either one of the values - e.g. [3] in AMPSYN - occurs very seldom, or because the two values might possibly not be told apart with great certainty by e.g. an automatic classification. At the same time, the values do not discriminate different **LA-types** such as e.g. the values [1] and [2] in F0SYN and F0PAR, cf. table 1.

1. **NUMBER** = Number of glottal pulses: [1] many periods (83.5%); [2] two to three periods (8.8%); [3] one period (7.3%); {2 3}
2. **DAMPING** = Special form of the damped wave: [1] relatively normal damping (42.4%); [2] strong exponential decay of the amplitude (2.6%); [3] "delta-like", triangular damping (24.4%); [4] "unusual" damping (30.1%); {2 4}
3. **AMPSYN** = Amplitude compared with left and right context (syntagmatic aspect): [1] normal (76.2%); [2] lower (23.3%); [3] higher (0.4%); {1 3}
4. **AMPPAR** = Amplitude variations inside the LA (paradigmatic aspect): [1] regular envelope, no variations (23.3%); [2] slightly irregular envelope (45.3%); [3] "diplophonic", i.e. regular variation between high and low amplitude (17.8%); [4] break down of envelope (12.7%); {1 2}
5. **F0SYN** = F0 compared with context (syntagmatic aspect): [1] regular, no variations (38.0%); [2] slightly irregular (15.1%); [3] subharmonic (25.2%); [4] extremely long period(s) or pause (20.3%); {1 2}
6. **F0PAR** = F0 variations inside the LA (paradigmatic aspect): [1] regular, no variations (39.7%); [2] slightly irregular (28.1%); [3] strong variations (25.3%); [4] periods not detectable (6.7%); {1 2}

The feature value [1] is always the default value as it is found regularly in non-la speech as well. A value was determined if it showed up during more of half of the la passage. A "compound type" (56 occurrences in the database) was determined if the la passage consisted of two or more clearly distinct parts that could be classified on their own. These parts were treated separately. In total 1251 LAs were labeled with MÜSLI.

RESULTS AND DISCUSSION

Out of all 1251 LAs 81% could be classified unequivocally and completely. 18% could also be classified, but with a disagreement in at least one feature value between the two phoneticians. In only 18 cases there was at least one feature value that could not be determined at all (feature value 0, cf. figure 6). The numbers given in the following always refer to all LAs except these 18 cases. For a grouping of the LAs into distinct LA-types, we first chose those combinations of feature values (sextuples) that occurred ≥ 10 times. These sextuples were grouped so that (near) default values were combined with as few as possible non-default values. We distinguish four different domains in the time signal: *Number*, *Damping*, *Amplitude*, and *Frequency*. In the following description, parentheses contain one or more of: 1. the relevant domains; 2. the number of the figure showing an example; 3. the terms used in [2] if they differ. Three LA-types could be differentiated with the help of **different domains**: GLOTTALIZATION (*Number* and *Frequency*, figure 1), DAMPING (*Damping*, figure 2, creak), DIPLOPHONIA (*Amplitude*, figure 3). Two LA-types could be differentiated **within one single domain**, namely SUBHARMONIC (figure 4, creak), and APERIODICITY (figure 5, creak or creaky voice) both having different values inside *Frequency* for F0SYN and F0PAR. In figure 6, the WASTE PAPER BASKET LA-type is illustrated with an example where two feature values (for AMPPAR and F0SYN) could not been defined. AMPSYN is no “distinctive feature” because it does not discriminate LA-types but it can characterize LAs in general. In figure 1–6 the sextuple of feature values is given in each caption in parentheses. Although a “standard” GLOTTALIZATION has only one period followed by a long pause, the example given in figure 1 represents roughly half of all the GLOTTALIZATIONs in our material.

Table 1: LA-*types and their characterization with MÜSLI*

LA-type (number of cases)	*Domains* & FEATURES					
	Number	*Damping*	*Amplitude*		*Frequency*	
	NUMBER	DAMPING	AMPSYN	AMPPAR	F0SYN	F0PAR
GLOTTALIZATION	’[23]’	3	1	[12]	’4’	[13]
(61/114/116)		[1234]	[123]			[123]
DAMPING	1	’[34]’	1	[12]	1	[12]
(161/292/680)	[123]	[234]	[123]	[124]	[12]	[12]
DIPLOPHONIA	1	1	1	’3’	[12]	[12]
(122/166/222)		[1234]	[123]			
SUBHARMONIC	1	[134]	1	[12]	’3’	’[12]’
(109/157/190)		[1234]	[123]	[124]		
APERIODICITY	1	[134]	[12]	2	[34]	’[34]’
(158/242/384)		[1234]	[123]	[1234]		

Table 1 shows the five LA-types and their characterization with special feature values. The columns can be interpreted as regular terms: between columns holds conjunction, within brackets holds disjunction. Combinatorically $3 \cdot 4 \cdot 3 \cdot 4 \cdot 4 \cdot 4 = 2304$ different sextuples can occur. In the first line of each LA-type the combinations are shown that entail ≥ 10 cases (narrow condition; 56 possible, 24 occurring sextuples). Weakening the conditions more cases can be classified; cf. the possible feature combinations in the second line of each LA-type (broad condition; 780 possible, 178 occurring sextuples). In the second line cells are left empty, whose terms do not differ from the corresponding terms in the first line. Cases that are comprised in line two are kept disjoint, i.e. there is no intersection of two LA-types. They represent so to speak **pure LA-types**. However, if we use as criterion only the “distinctive feature” values quoted in line one, i.e. for the other features all values are valid (very broad condition), we get 3552 possible and 247 occurring sextuples. 532 cases belong to more than one LA-type, 83% of them forming an intersection of DAMPING with other

LA-types. In the first column of table 1 the number of cases for narrow/broad/very broad conditions are given in parentheses below the name of each LA-type.

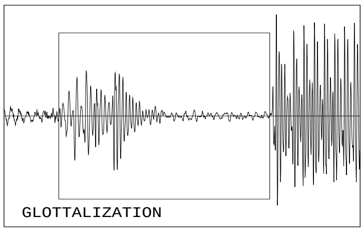

Figure 1 (231243)

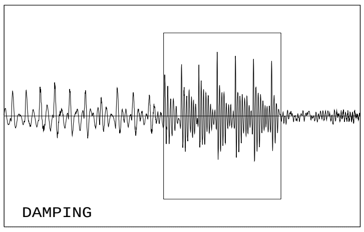

Figure 2 (141211)

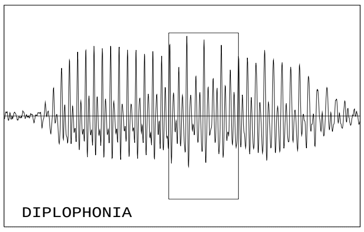

Figure 3 (111311)

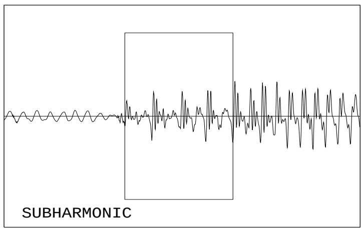

Figure 4 (111231)

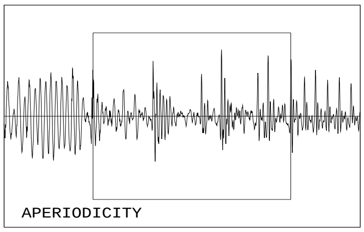

Figure 5 (131243)

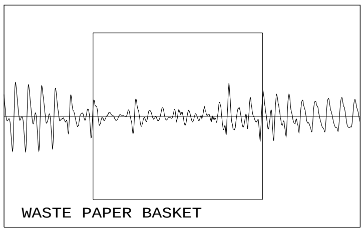

Figure 6 (111004)

FINAL REMARKS

It can be doubted that the features are distinctive phonologically but at least some of them might constitute allophones occurring in different contexts, while others might describe simply free variants. Yet, to our knowledge the feature matrix in table 1 is the first attempt to describe a large corpus of LAs systematically and exhaustively with a feature approach; it seems to work reasonably well. MÜSLI should, however, not be taken as **the** final classification scheme for LAs but rather as a starting point for further investigation. Other possible features as e.g. spectral tilt, breathiness or (partial) devoicing could be taken into consideration as well. The next step will be the automatic extraction of the different features and then hopefully a more straightforward but at the same time more robust feature description and a reduction of overspecification. It should be investigated further whether different LA-types can be discriminated perceptually, whether different LA-types have different functions such as e.g. boundary marking, and if the different LA-types are speaker-, language-, or register-specific.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the authors are responsible for the contents of this paper.

References

[1] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. *Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody.* In this volume.

[2] D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation.* PhD thesis, Chalmers University, Göteborg/Lund, 1988.

[3] A. Kießling, R. Kompe, E. Nöth, and A. Batliner. *Irregularitäten im Sprachsignal — störend oder informativ?* In R. Hoffmann, editor, *Elektronische Signalverarbeitung*, volume 8 of *Studientexte zur Sprachkommunikation*, pages 104–108. TU Dresden, 1991.