Proceedings of the International Workshop on Paralinguistic Speech - between Models and Data (ParaLing'07), Saarbrücken, August 3, 2007. eds.: Marc Schröder, Anton Batliner, Christophe d'Alessandro, pp. 17-22.

Laryngealizations and Emotions: How Many Babushkas?

Anton Batliner, Stefan Steidl, and Elmar Nöth

Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
batliner@informatik.uni-erlangen.de

Abstract

It has been claimed that voice quality traits including irregular phonation such as creaky voice (laryngealization) serve several functions, amongst them being the marking of emotions; accordingly, they should be used for automatic recognition of these phenomena. However, laryngealizations marking emotional states have mostly been found for acted or synthesized data. First results using real-life data do not corroborate such an impact of laryngealized speech for the marking of emotions. For a German speech database with realistic emotions, we corrected manually the automatically extracted F0 values and operationalised extraction errors as indications of laryngealized passages. Even if at first sight, it seems plausible that some emotions might display higher frequencies of laryngealizations, at a closer look, we find that it is rather a combination of speaker-specific traits and lexical/segmental characteristics which causes the specific distribution. We argue that the multifunctionality of phenomena such as voice quality traits makes it rather difficult to transfer results from acted/synthesized data onto realistic speech data, and especially, to employ them for speaker-independent automatic processing, as long as very large databases modelling diversity to a much higher extent are not available.

Index Terms: laryngealizations, emotions, speaker dependency, segmental structure

1. Introduction

It is well-known that phonetic parameters such as pitch, energy, duration, or spectrum are multi-functional, i.e. they serve different purposes. Let us take pitch as example: In the high days of intonation models, pitch was held responsible for the marking of boundaries, word- and sentence accents, emphasis, salience, emotions, etc. During the last years, however, it has been shown that F0 is of less importance for the marking of accentuation, in relation to other parameters such as energy and duration, cf. [1, 2] and [17]. The same might be true for emotion recognition; again, we do not know yet whether this might be due to pitch simply being less important, or to a combination of extraction errors, speaker specific traits and other factors which all are difficult to model, esp. with sparse data.

The findings that parameters such as pitch are not that much important in general went along with a shift of interest, away from laboratory studies with tightly controlled (real or synthesised) speech with identical segmental structure towards more realistic, spontaneous speech. Contrasting *'lab speech'* with *'real speech'*, on the one hand we want to stress that we are more interested in *real speech* — simply because it has to be studied in order to get high performance for automatic speech processing systems. On the other hand, both approaches are only approximations towards a full understanding of modelling human perception and comprehension: humans are definitely able to use, so to speak, the magnifying glass for looking at and processing speech phenomena — in analogy to experiments with *lab speech*. But humans normally do not fully and deeply process single parameters; they use shallow processing, taking into account a plethora of knowledge sources, especially linguistic information which is of course lacking if segmental structure is identical. Thus using simply larger speech databases for training the classifier — this is state-of-the-art nowadays — of course falls short of the broad range of human capacities.

The same might hold for voice quality parameters such as breathiness, harshness, or irregular phonation. Note that we will not deal directly with acoustic parameters such as spectral tilt, zero-crossing-rate, etc. which characterize voice quality; we only have a look at one complex phenomenon which will be introduced in the next section 2.

2. Laryngealizations: the Phenomenon and its Functions

The normal speech register 'modal voice' comprises an F0 range from about 60 to 250 Hz for male speakers and an 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 not much investigated until now. There is a variety of different terms for this phenomenon which are used more or less synonymously: irregular phonation, 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.

A thorough account of voice quality is given in [20]. In [3], five different types of LAs have been established: glottalization, diplophonia, damping, sub-harmonic, and aperiodicity, cf. Fig. 1. Table 1 displays different functions of LAs which can be linguistic or paralinguistic. They can be caused either by higher effort or by relaxation; in the first case, they go together with *accentuation* (prominence) which is, of course, a *local* phenomenon. (Actually, it might be that LA is not denoting accentuation but can be accompanied by it, if the vowel is 'LA-prone', cf. below; however, in such cases, LA cannot be caused by relaxation.) A typical place for relaxation is *the end of an utterance*; by that, *turn-taking* can be signalled to the dialogue partner; this is again a *local* phenomenon: [22] report that different types of LAs are used in (British and American) English conversations for holding the floor (filled pauses with glottal closure, no evidence of creaky phonation) and for yielding the floor (filled pauses with lax creaky phonation, no glottal closure). *Word boundaries* in the hiatus, i.e. word final vowel followed by word initial vowel, can be marked by LAs. Boundary marking which is, of course, *local*, with such

Proceedings of the International Workshop on Paralinguistic Speech - between Models and Data (ParaLing'07), Saarbrücken, August 3, 2007. eds.: Marc Schröder, Anton Batliner, Christophe d'Alessandro, pp. 17-22.

irregular phonation is dealt with in [13], [18], and [23]. It is well known that back *vowels* such as *[a]* tend to be more laryngealized than front vowels such as *[i]* (*local* phenomenon). A language-specific use of LAs can be either due to phonotactics, as in German, where every vowel in word-initial position is 'glottalized', or phonemes can be creaky, cf. [19]; this is a *local* phenomenon, denoting the *native language*. Normally, specific segments which are laryngealized characterize languages, cf. for vowels [11]; the Danish glottal catch (stød) [9] can be found in vowels and consonants.

[21, p. 194ff.] lists different uses and functions of 'creak' phonation, amongst them the paralinguistic function 'bored resignation' in English RP, 'commiseration and complaint' in Tzeltal, and 'apology or supplication' in an Otomanguean language of Central America. Extra- and paralinguistically, LAs can be a marker of personal identity and/or social class; normally, LAs are a marker of higher class speech. [26] quote evidence that not only for human voices but for mammals in general, 'non-linear phenomena' (i.e. irregular phonation/LA) can denote individuality and status (pitch as an indicator of a large body size and/or social dominance; *"... subharmonic components might be used to mimic a low-sounding voice"*).

Note that all these characteristics which per se are **not** characteristics of single speakers can be — maybe apart from the language-specific phonemes — used more or less distinctly by different speakers. As for the para- and extra-linguistic function of LAs, speakers can simply use them throughout to a higher extent; such *speaker idiosyncrasies* are *local - global*. 'Creaky superstars' like Tom Waits are well-known. The reason might be unknown, or due to one or more of the following factors: *speaker pathology* (*global*), *too many drinks/cigarettes* (*temporary*), *competence/power* (*global / temporary*), or *social class membership* (*local/global/temporary*).

Emotional states such as *despair, boredom, sadness*, etc. are *short-term* or *temporary*. Bad news are communicated with breathy and creaky voice [10], boredom with lax creaky voice, and to a smaller extent, sadness with creaky voice [12]. [7] report for perception experiments with synthesized stimuli that disgust is conveyed with creaky voice. [8] found, for one female Japanese speaker, creaky voice in imitated sadness but not in spontaneous sadness; thus they assume a social connotation of creaky voice. To display boredom or to display upper-class behaviour might co-incide; the same can happen if someone who permanently uses LAs as speaker-specific trait, speaks about a sad story. On the other hand, at first sight, speakers who exhibit LAs as an idiosyncratic trait can make a sad impression without actually being sad. A common denominator for some of the paralinguistic functions of LAs might be inactivity/passivity in mood (boredom, sadness, etc.) corresponding to relaxation — which is one of the possible physiological source of LAs. However, this is no must: if LAs are used to signal competence/power, then the basic attitude need not be passivity.

Further functions of LAs are reported in [15]. There are only a few studies dealing with the automatic detection of LAs, cf. [16, 14].

The caveat has to be made that we are speaking of a sort of 'cover phenomenon' covering different sub-phenomena and different temporal traits: some are very short and might rather be perceived as segmental features, i.e. not as supra-segmental, prosodic features that are sort of modulated onto the speech wave. Of course, there are prototypical cases — no LA at all and laryngealized throughout — which easily can be told apart. But we simply do not know yet when people will produce which amount of LA and how an automatic classifier can model it.

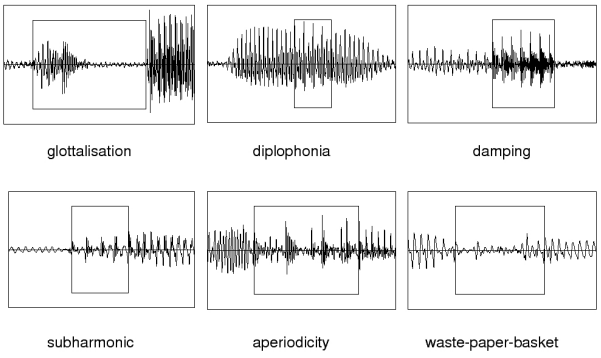

Figure 1: Five different types of LAs, in prototypical examples, and the wast-paper-basket type; from [3]

Table 1: Different Functions of Laryngealizations

phenomenon	**domain**
linguistic functions: phonotactics, grammar, ...	
accentuation	local
vowels	local
word boundaries	local
native language	local
the end of an utterance	local
paralinguistic functions: speaker characteristics	
speaker idiosyncrasies	local - global
speaker pathology	global
too many drinks / cigarettes	temporary
competence / power	global / temporary
social class membership	local / global / temporary
emotional states	short-term or temporary
explanation of 'domains'	
local: phonotactically definable (utterance-final, word-initial, etc.) or phone-dependent	
global: persistent	
short-term: definable on the time axis in sec./min.	
temporary: longer than short-term but not global	

3. Material and Annotation

The database used for this study is a German corpus with recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [5, 4] and other papers quoted therein. The speech is spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend. They were led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator who causes the AIBO to perform a fixed, predetermined sequence of actions; sometimes the AIBO behaved disobediently, by that provoking emotional reactions. The data was collected at two different schools from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). The recordings were segmented automatically into 'turns' using a pause threshold of 1500 msec. Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated independently from each other each word as *neutral* (default) or as belonging to one of ten other classes. If three or

Proceedings of the International Workshop on Paralinguistic Speech - between Models and Data (ParaLing'07), Saarbrücken, August 3, 2007. eds.: Marc Schröder, Anton Batliner, Christophe d'Alessandro, pp. 17-22.

more labelers agree, the label is attributed to the word (majority voting MV); in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), **E***mphatic* (2528), *helpless* (3), **T***ouchy*, i.e., irritated (225), **A***ngry* (84), **M***otherese* (1260), *bored* (11), **R***eprimanding* (310), *rest*, i.e. non-neutral, but not belonging to the other categories (3), **N***eutral* (39169). 4707 words had no MV (**U***ndecided*); all in all, there were 48401 words. We will only deal with labels with more than 50 tokens (first character given bold-faced and recte); as the figures for *joyful* are very low for the different constellations to be addressed, this state will not be dealt with in the following.

Some of the labels are very sparse. Therefore, *neutral* and *emphatic* items were down-sampled, and *touchy* and *reprimanding*, together with *angry*, were mapped onto **A***ngry* as representing different but closely related kinds of negative attitude. This more balanced 4-class problem consists of 1557 words for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), 1645 words for **E***mphatic* (**E**), and 1645 for **N***eutral* (**N**). Cases where less than three labelers agreed were omitted as well as those cases where other than these four main classes were labeled. Interlabeler correspondence is dealt with in [24].

In this paper, we deal with those turns which at least contain one word attributed to one of these four cover classes; this amounts to 17.618 words in total. However, we will go back to the more fine-grained labels, i.e., we do not want to deal with the cover-class **A***ngry* but with the detailed classes *touchy*, i.e., irritated, *angry*, and *reprimanding*.

Table 2: Gross F0 errors (>10 % deviation) within words (613 278 frames, 67,9 % voiced.)

type	# frames	percent
identical	574 485	93.67
small errors	452	0.07
voiced errors	8 804	1.43
unvoiced errors	1 877	0.30
octave errors ↓	23 498	3.83
octave errors ↑	239	0.03
other gross errors	3 923	0.63

Our database might seem to be untypical because it deals with children's speech; however, children represent just one of the usual partitions of the world's population into sub-groups such as women/men, upper/lower class, or different nations. Of course, automatic procedures have to adapt to this specific group — an Automatic Speech Recognition (ASR) system trained with adult speech initially performs poorly with children's speech. For instance, F0 range for children is different from the one for adults: upper bound should be higher than the 550 Hz normally assumed for female voices; in our experience, 700 Hz could be a reasonable upper bound. This is nothing special but just the normal necessity to adapt to specific sub-groups. So far, we have found no indication that these children behave differently in a principled way, as far as speech in general or emotional states conveyed via speech are concerned. Moreover, the database is typical for realistic, spontaneous (neutral and) emotional speech: we are faced with the well-known sparse data problem — neutral is by far the most frequent class. The linguistic structure of the children's utterances is not too uniform, as it might be if it is only pure, short commands; on the other hand, it displays specific traits, for instance, many vocatives because these are representative for direct addressing by giving commands.

4. Manual Pitch Correction

Word segmentation based on forced alignment was corrected manually; for automatic F0 extraction, we chose the ESPS algorithm [25] because it is well established, a software is freely available, and it is often used for benchmarking. The pitch values for all words within all turns which contained at least one of the AMEN words described above were corrected manually by the first author. Actually, a better name instead of 'corrected' would be something like 'smoothed and adjusted to human perception'. The basic idea behind is that those irregularities which are called *creak/creaky voice/laryngealizations/...* are modulated onto the pitch contour and not perceived as jumps up or down [3]. Table 2 displays percentage of corrected F0 values and their types. It can be seen that some 6.7% of all voiced frames in the words were corrected as for octave or other gross errors. The correction dealt mostly with the following phenomena:

(1) **octave jumps** – correction by one octave up, in some rare cases two octaves up or one octave down. This concerns rather smooth curves which had to be transposed. In most cases, it is a matter of irregular phonation; in such cases, the extraction algorithm modelled pitch rather 'close to the signal', not 'close to perception'. In a few cases, however, no clear sign of laryngealization can be observed. Sometimes, the context and/or perception had to decide whether an octave jump has to be corrected or not. If the whole word is laryngealized and the impression is low pitch throughout, then laryngealization is not modulated onto pitch; in these cases, no octave jump was corrected.

(2) **smoothing at irregularities** – normally laryngealizations or voiceless parts which wrongly have been classified as voiced: the ESPS curve is not smoothed but irregular; here, often the context to the left and to the right was interpolated in order to result in a smoothed curve; in the case of voiceless parts, F0 was set to 0.

(3) **other phenomena**, for instance irregularities at transitions which not necessarily are due to irregular phonation. **smoothing at transitions** is admittedly a bit touchy — when should it be done if the phenomenon is well known, e.g. in the case of higher F0 values after voiceless consonants. Sometimes, the context and/or perception had to decide whether an octave jump has to be corrected or not. A typical problem is a hiatus, e.g. the sequence of one word, ending in a vowel, followed by /Aibo/. Here, everything can happen: the perception is rather no pitch movement but 'something' modulated onto the pitch curve; F0 values can be voiceless, i.e. zero, or we can observe an octave jump down, an octave jump up, fully irregular F0 values, or values from low to higher. Here, we sometimes interpolated, sometimes used 'double F0', sometimes we did not correct (in the case of 'from low to higher'). Sometimes we could not find clear criteria for the one or the other solution, at least not without too much effort. In VCV sequences within a word, e.g., "Aibo", the plosive sometimes was set to voiceless even if voiced would have been possible - F0 postprocessing sometimes interpolates in such cases anyway. In some rare cases, it had to be educated guessing and not really based on strong criteria.

Figures 2 to 3 show an example turn (first and second word) with F0 corrections: below, the time signal, in the middle, the spectrum, and above, F0 values given per frame à 10 ms. Corrected F0 values (in red) are displayed with a grey background; printed black & white, these corrected values are light grey. The first part (the [a] in [aI]) of /Aibo/ is clearly laryngealized, first glottalization, then diplophonia, and in the last irregular part, aperiodicity, following the terminology in [3]; an unvoiced de-

Proceedings of the International Workshop on Paralinguistic Speech - between Models and Data (ParaLing'07), Saarbrücken, August 3, 2007. eds.: Marc Schröder, Anton Batliner, Christophe d'Alessandro, pp. 17-22.

cision was made for the intervocalic plosive [b] (note that this is regular in south German dialects). The [a] in /tanz/, however, diplays no clear sign of irregular phonation. It simply would have been too much effort to analyse in detail whether signs of laryngealization can be detected in the 'magnifying glass' or not. In both cases, we used the 'double F0 value' function of the labelling tool.[1]

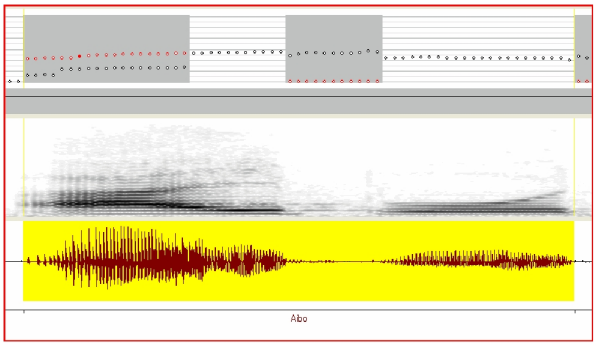

Figure 2: Example: Pitch correction at laryngealization; first part of the turn "Aibo, tanz!': [aIbo]

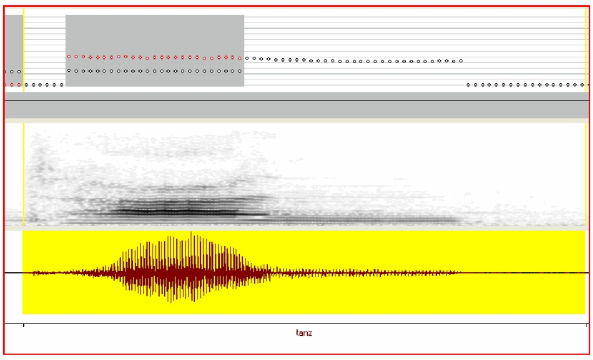

Figure 3: Example: Pitch correction at laryngealization; second part of the turn "Aibo, tanz!': [tants]

In this study, we are not interested in the specific German phenomenon glottalization at a word-initial vowel; neither are we interested in small irregular perturbations at transitions. We decided in favour of a heuristic threshold of three frames which corresponds, for example, to six periods at 200 Hz; if the irregular passage consisting of gross F0 errors (>10 % deviation) is longer than three frames, then we assume an LA. By that, we so to speak operationalize our pitch correction for identifying LAs. As mentioned above, we have to tolerate a certain amount of 'white noise', i.e. extraction errors where no clear sign of LAs could be seen at first sight. A precise figure is not possible because this would have demanded exactly an effort greater by some order of magnitude — which we had to avoid; we estimate the percentage of such cases being below 10%.

Table 3: Distribution of emotions and LAs for 17618 words in percent

words	*motherese*	*neutral*	*undecid.*	*emphatic*	*touchy*	*reprim.*	*angry*
all	4.5	68.3	7.5	14.2	2.2	2.4	0.7
with LA	12.3	12.1	16.0	14.0	21.0	22.9	7.1
not[aI,aU]	4.9	70.3	5.6	16.5	1.6	0.5	0.4
[aI,aU]	3.2	61.3	13.1	9.6	3.8	7.6	1.4

5. Distribution of Sub-sets: Unveiling the Babushkas

For a complete account of our data, we mapped all those words which do not belong to one of the MV classes, onto the class *undecided*. Table 3 displays percent of emotions in the database (summing up to 100% modulo rounding errors, first line), and of these emotions, percentage laryngealized (second line, not summing up to 100%), percent words without the diphthongs [aI] or [aU] (third line), and words with the diphthongs [aI] or [aU] (fourth line; third and fourth line again summing up to 100% modulo rounding errors). In the second line it can be seen that the proportion of LAs per emotion are higher for *touchy* and *reprimanding* and lower for *angry* than for the rest. We can try an explanation: negative attitudes such as *touchy* and *reprimanding* imply a sort of superiority; this holds for *angry* as well but in this case, higher arousal might contradict the relaxation which normally is a prerequisite of LAs.

Now we start unveiling the Babushkas (Babushka dolls is a set of nested dolls of decreasing sizes placed one inside another). The first Babushka comes up if we have a look at lines three and four: relatively, words with [aI] or [aU] are much more frequent for the negative emotions than for the other ones. Note that for instance, the word /Aibo/ — almost always used as vocative — has 2769 tokens in our subset which amounts to 60.8% of all words with the diphthongs [aI] or [aU]. As there are 17618 words in total in our subset, the word /Aibo/ amounts to 15.7% of all words.

The next Babushka can be seen in Fig. 4: obviously LAs per speaker are very unbalanced. With the 'scree'-criterion (when a steep slope levels off into a shallow slope) we can tell apart the five speakers with a high percentage of LAs (the five speakers to the right in Fig. 4) from the rest.

The prevalence of words with [aI,aU] for the two emotions *touchy* and esp. *reprimanding* can be seen in Fig. 5. In Figures 6 to 8, we split into the two speaker groups, the one with not that many LAs, the other one consisting of the five speakers which heavily make use of LAs, cf. Fig. 4. Fig. 6 displays all words; in Fig. 7 and Fig. 8, a break down into words without, and into words with [aI,aU] is shown. Even for words without [aI,aU], a pronounced tendency towards using more LAs is shown for the heavy LA speakers.

By unveiling the Babushkas we have seen that LAs are mostly 'caused' by segmental structure and by speaker idiosyncrasies. But even if we control these factors in Fig. 6 to 8, by displaying different sub-groups, there is still a tendency for *touchy* and *reprimanding* to show a higher proportion of LAs and for *angry*, to show a lower proportion of LAs. But here

[1]The tool **eLabel** has been developed by the second author and is freely available at: *http://www5.cs.fau.de/humaine/download/*.

Proceedings of the International Workshop on Paralinguistic Speech - between Models and Data (ParaLing'07), Saarbrücken, August 3, 2007. eds.: Marc Schröder, Anton Batliner, Christophe d'Alessandro, pp. 17-22.

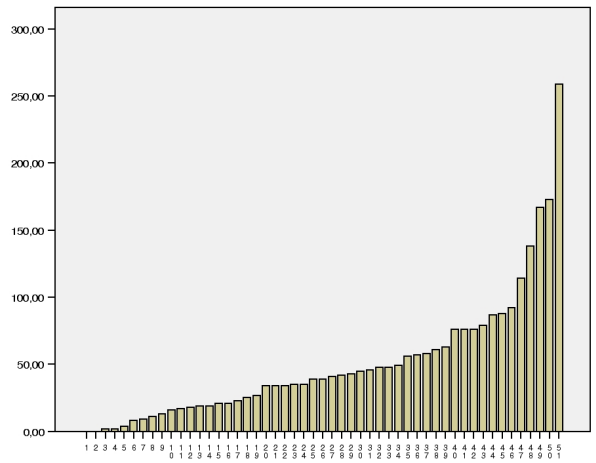

Figure 4: # of laryngealizations per speaker

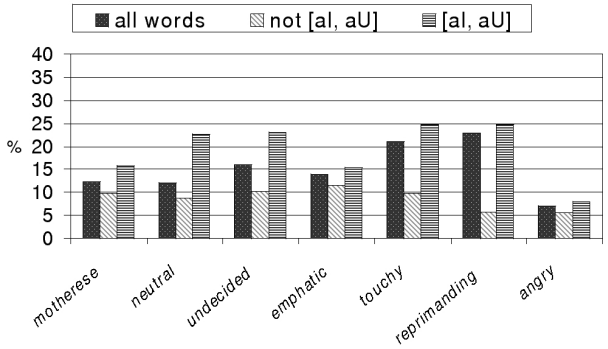

Figure 5: All speakers, percent laryngealized per emotion: all words, not[aI,aU], and [aI,aU]

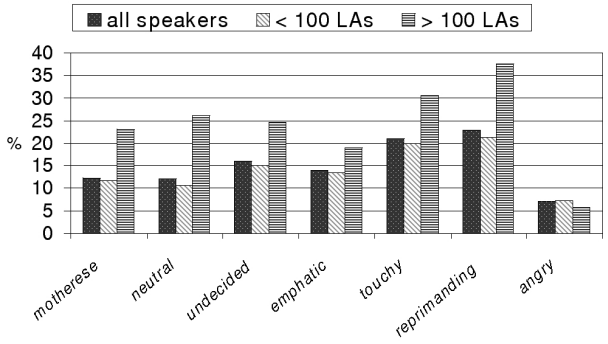

Figure 6: All words, percent laryngealized per emotion: all speakers, speakers with $<$ 100 LAs, speakers with $>$ 100 LAs

again we are faced with the sparse data problem: for instance, the frequency of *angry* is generally very low, and the one for *touchy* and *reprimanding* is not much higher, either; thus a meaningful statistical statement is not possible.

6. Discussion and Concluding Remarks

It seems that at least in our data, LAs 'induced' by speaker-idiosyncrasies and/or segmental structure make up the largest

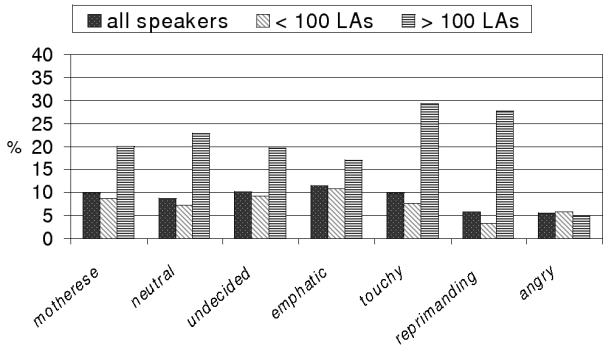

Figure 7: All words without [aI,aU], percent laryngealized per emotion: all speakers, speakers with $<$ 100 LAs, speakers with $>$ 100 LAs

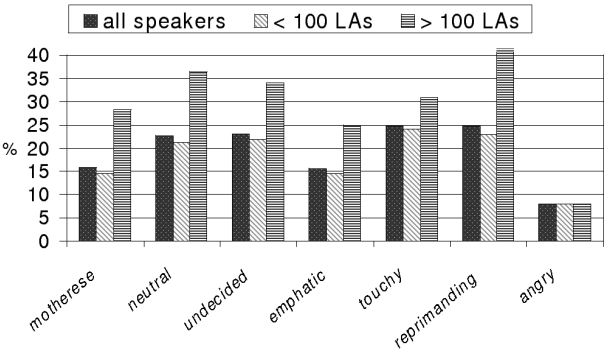

Figure 8: All words with [aI,aU], percent laryngealized per emotion: all speakers, speakers with $<$ 100 LAs, speakers with $>$ 100 LAs

part of laryngealized, emotional words. This does not mean to exclude the explanation given above, that negative attitudes might go along more often with LAs than positive ones. Obviously, if such a tendency exists, it is not strong enough to help in automatic recognition: we only got results close to chance level when we tried to use LAs as such within automatic classification. For an extensive classification study assessing different types of features as for their relevance, we found out that voice quality features are the least important ones: performance at about 50% for four classes, best types being duration and energy at about 60%. Note that this definitely does not mean that voice quality features in general or LAs in particular are not relevant at all. It might be, however, that the factors we have been dealing with (esp. speaker-idiosyncrasy and segment structure) so to speak smear the contribution of LAs within a speaker-independent study.

We did not prove either that LAs never signal some other emotional states, because in our data, emotions such as *sadness* (cf. the database processed in [6]) or *boredom* were not found. However, we want to stress that in our opinion, acted or synthesised data — which so far represent the bulk of evidence on voice quality features in emotions — can be used as sort of 'heuristic inspiration' but must not simply be transferred generically onto realistic data. As [8, p. 20] put it: 'In acted emotion, the speaker is volitionally changing the acoustic signal to impart to the listener a mental or emotional state (paralanguage) while in spontaneous emotion the speaker is working at main-

taining the acoustic signal to convey the intended message even through emotional interruptions (nonlanguage).'

It could be argued that laryngeal control does not fully mature until young adulthood; therefore, the choice of our 10 to 12 year old children would prevent a clearer pattern from emerging. However, although children's data might be more 'noisy' than adult's data, there is no indication that the general trends would change: everything that could be observed at our children's speech can be found in adult data — which should, of course, be used systematically for cross-checking our findings.

To sum up, we could illustrate the multi-functionality and speaker-dependency of LAs; thus it might be less likely that they are very useful as a generic feature within emotion classification. This can of course be different in a personalized setting, or with training databases which are, by some order of magnitude, larger than those available today.

7. Acknowledgements

This work was partly funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The responsibility lies with the authors.

8. References

[1] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. of the 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, 1999.

[2] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *In Proc. 7th Eurospeech*, pages 2781–2784, Aalborg, 2001.

[3] A. Batliner, S. Burger, B. Johne, and A. Kießling. MÜSLI: A Classification Scheme For Laryngealizations. In D. House and P. Touati, editors, *Proc. of an ESCA Workshop on Prosody*, pages 176–179. Lund University, Department of Linguistics, Lund, Sept. 1993.

[4] A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private Emotions vs. Social Interaction — a Data-Driven Approach towards Analysing Emotion in Speech. *UMUAI — User Modelling and User-Adapted Interaction*, 2007. to appear.

[5] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana, 2006.

[6] L. Devillers and L. Vidrascu. Real-life Emotion Recognition in Speech. In C. Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg - Berlin - New York, 2007. to appear.

[7] C. Drioli, G. Tisato, P. Cosi, and F. Tesser. Emotions and Voice Quality: Experiments with Sinusoidal Modeling. In *Proceedings of VOQUAL'03*, pages 127–132, Geneva, 2003.

[8] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya. Exploratory Study of Some Acoustic and Articulatory Characteristics of *Sad* Speech. *Phonetica*, 63:1–25, 2004.

[9] E. Fischer-Jørgensen. Phonetic analysis of the stød in standard Danish. *Phonetica*, 46:1–59, 1989.

[10] J. Freese and D. W. Maynard. Prosodic features of bad news and good news in conversation. *Language in Society*, 27:195–219, 1998.

[11] C. Gerfen and K. Baker. The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, pages 311–334, 2005.

[12] C. Gobl and A. Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2):189–212, 2003.

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

[14] C. Ishi, H. Ishiguro, and N. Hagita. Proposal of Acoustic Measures for Automatic Detection of Vocal Fry. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 481–484, Lisbon, 2005.

[15] C. Ishi, H. Ishiguro, and N. Hagita. Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction. In *Proc. of Speech Prosody 2006*, pages 883–886, Dresden, 2006.

[16] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Voice Source State as a Source of Information in Speech Recognition: Detection of Laryngealizations. In A. Rubio Ayuso and J. López Soler, editors, *Speech Recognition and Coding. New Advances and Trends*, volume 147 of *NATO ASI Series F*, pages 329–332. Springer, Berlin, 1995.

[17] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts Prominence; Fundamental Frequency lends little. *Journal of Acoustical Society of America*, 11:1038–1054, 2005.

[18] S. Kushan and J. Slifka. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In *Proc. of Speech Prosody 2006*, pages 795–798, Dresden, 2006.

[19] P. Ladefoged and I. Maddieson. *The Sound of the World's Languages*. Blackwell, Oxford, 1996.

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

[21] J. Laver. *Principles of Phonetics*. Cambridge University Press, Cambridge, 1994.

[22] J. Local and J. Kelly. Projection and 'silences': notes on phonetic and conversational structure. *Human Studies*, 9:185–204, 1986.

[23] A. Ní Chasaide and C. Gobl. Voice Quality and f_0 in Prosody: Towards a Holistic Account. In *Proc. of Speech Prosody 2004*, Nara, Japan, 2004. 4 pages, no pagination.

[24] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. of ICASSP 2005*, pages 317–320, Philadelphia, 2005.

[25] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech coding and synthesis*, pages 495–518. Elsevier Science, 1995.

[26] I. Wilden, H. Herzel, G. Peters, and G. Tembrock. Subharmonics, biphonation, and deterministic chaos in mammal vocalization. *Bioacoustics*, 9:171–196, 1998.