THE IMPACT OF F0 EXTRACTION ERRORS ON THE CLASSIFICATION OF PROMINENCE AND EMOTION

A. Batliner[1], *S. Steidl*[1], *B. Schuller*[2], *D. Seppi*[3], *T. Vogt*[4], *L. Devillers*[5], *L. Vidrascu*[5], *N. Amir*[6], *L. Kessous*[6], *V. Aharonson*[7]

[1] Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Institute for Human-Machine Communication, Technische Universität München, Germany
[3] Fondazione Bruno Kessler – irst, Trento, Italy
[4] Multimedia Concepts and their Applications, University of Augsburg, Germany
[5] Spoken Language Processing Group, LIMSI-CNRS, Orsay Cedex, France
[6] Dep. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel
[7] Tel Aviv Academic College of Engineering, Tel Aviv, Israel
batliner@informatik.uni-erlangen.de

ABSTRACT

Traditionally, it has been assumed that pitch is the most important prosodic feature for the marking of prominence, and of other phenomena such as the marking of boundaries or emotions. This role has been put into question by recent studies. As nowadays larger databases are always being processed automatically, it is not clear up to what extent the possibly lower relevance of pitch can be attributed to extraction errors or to other factors. We present some ideas as for a phenomenological difference between pitch and duration, and compare the performance of automatically extracted F0 values and of manually corrected F0 values for the automatic recognition of prominence and emotion in spontaneous speech (children giving commands to a pet robot). The difference in classification performance between corrected and automatically extracted pitch features turns out to be consistent but not very pronounced.

Keywords: pitch, automatic extraction, manual correction, automatic classification

1. INTRODUCTION

Amongst the 'traditional' prosodic parameters pitch, duration, and energy, pitch has been attributed a central role for signalling linguistic as well as paralinguistic phenomena such as syntactic-prosodic boundaries, accentuation/prominence, and emotional/affective states – just to mention some of the most important ones. Especially for linguistic phenomena, this tendency showed up in several phonological intonation models whereas for paralinguistic phenomena, such models have been used less often (but cf. [6]); instead, more parametric models — be this explicit intonation models or a straightforward use of pitch values — have been employed. The shift from closely controlled laboratory speech to less restricted – even realistic, spontaneous – speech as the object of investigation during the last decade is mirrored in the use of a plethora of automatically extracted acoustic features and automatic classification procedures instead of a few characterising features and inferential statistics such as analysis of variance. Although in some of these studies, the relevance of single features and/or feature groups has been addressed, we are far from a clear picture of the relative importance of their impact on the classification tasks. This is due to at least three factors: first, universally valid statements cannot be made based on only one or only a few specific databases; second, the availabilty of extraction algorithms and 'brute force' computation methods generating hundreds or even thousands of features makes it rather difficult to point out the 'most important ones'; third, automatic feature extraction is always error-prone. This holds especially for pitch – there is no error-free pitch extraction algorithm. In recent years, it has been claimed that F0 is of minor importance in relation to other parameters such as energy and duration, for the marking of boundaries and accents in German and English [1] (automatically extracted pitch values), and for the marking of prominence in English [5] (manually corrected pitch values). In the present paper, we will address the third factor: the impact of erroneous pitch extraction on the automatic classification of prominence and emotional states in spontaneous speech. In this paper, we do not distinguish between the acoustic phenomenon F0 and the perceptual phenomenon pitch; actually many of our features are normalized (as to mean values or logarithmically) and thus more closely related to perception than the raw F0 values.

2. MATERIAL AND ANNOTATION

The database used is a German corpus with recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [3] and other papers quoted therein. The children were led to believe that the AIBO was responding to their commands, whereas the robot was actually being controlled by a human operator who caused the AIBO to perform a fixed, predetermined sequence of actions; sometimes the AIBO behaved disobediently, thereby 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). Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantization 16bit, down-sampled to 16 kHz). 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 each word independently from each other as neutral (default) or as belonging to one of ten other classes. If three or more labellers agreed, the label was attributed to the word (majority voting MV). All in all, there were 48401 words. As some of the labels are very sparse, they were down-sampled or mapped onto cover classes [3]. This resulted in a more balanced 4-class problem consisting of 1557 words for **A***ngry* as representing different but closely related kinds of negative attitude, 1224 words for **M***otherese*, 1645 words for **E***mphatic*, and 1645 for **N***eutral*.

Eventually, the word-based labels were mapped onto so-called chunk-based labels, again with an MV decision similar to the one described in [3], because semantically (and by that, syntactically) meaningful chunks most probably can better be mapped onto 'emotional units' than turns containing up to $>$50 words. We performed manually a coarse syntactic labelling with the following chunk triggering boundaries: at main clauses, free phrases, and between adjacent /Aibo/ instances because repetitions of vocatives make emotional coloring more likely. Spontaneous speech, especially in such scenarios as 'giving commands to a pet (robot)', is quite often not well-formed syntactically: no clear structural indication is found, let alone interpunction. We therefore used a prosodic criterion in addition: if the pause between words is $\geq$500 msec, we assume a chunk boundary. The length of the pauses between words was obtained from the manually corrected word segmentation. Based on 3996 turns, this procedure yielded a subset of 4543 chunks (914 **A***ngry*, 586 **M***otherese*, 1045 **E***mphatic*, and 1998 **N***eutral*) with an average length of 2.9 words per chunk.

In this paper, we will address the 4-class problem described above and, in addition, the 2-class sub-problem **E***mphatic* vs. **N***eutral*; by that, we can 'simulate' the more traditional problem [$\pm$ PROMINENCE]. This linguistic phenomenon is of course not the same as the type of emphasis found here, serving as a sort of 'pre-stage' to negative emotion in emotional speech; however, the both phenomena have quite a lot in common although we are not aware of any study that systematically has investigated their relationship. Content words, e.g., are more prone to be linguistically prominent and to be marked 'emotionally' than function words. [4] has shown that "all except for 6%" of accentuated words co-incide with words that were labelled as emphatic in an emotional annotation. Of course, this result cannot simply be transferred onto other data but we can assume that both phenomena are very much alike — albeit not identical.

3. MANUAL PITCH CORRECTION

Word segmentation based on forced alignment was corrected manually; for frame-based automatic (*aut*) F0 extraction, we chose the ESPS algorithm [7] because it is well established, a software is freely available, and it is often used for benchmarking. The pitch values of the 3996 turns were corrected manually (*corr*) by the first author. Actually, a better term 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/laryngealisations/...* are modulated onto the pitch contour and not perceived as jumps up or down [2]. Thus, the correction dealt mostly with the following phenomena:

octave jumps – correction by one octave up, in some rare cases two octaves up or one octave down. This usually involved 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 'close to the signal' rather than 'close to perception'. In a few cases, however, no clear sign of laryngealisation could be observed. Sometimes, the context and/or perception had to determine whether an octave jump had to be corrected or not. If the whole word is laryngealised and the impression is low F0 throughout, then laryngealisation is not modulated onto pitch; in these cases, no octave jump was corrected.

smoothing at irregularities – normally at laryngealisations or voiceless parts which were wrongly classified as voiced. The ESPS curve is not smoothed but irregular; here, often the context to the

Table 1: 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

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.

Table 1 displays percentage of corrected F0 values and their types. It can be seen that some 6% of all voiced frames displayed octave or other gross errors.

4. WHAT IS PITCH?

There is an obvious difference between pitch and duration: the latter is *one-dimensional* — longer or shorter on the time- (x-) axis. Even if, under certain circumstances, short duration can encode prominence, most often, it is the other way round. (Note that we are speaking here of 'prominence' in a broad meaning, not only of prominence denoting stress/accentuation.) F0, however, behaves differently: it is not only high vs. low pitch, it is the whole configuration, i.e. specific tunes, which are prominent. For emotion encoding, the same might hold as for accent encoding: in the tone sequence terminology, accents can be marked by L*H or H*L, i.e. by two 'opposite' configurations, whereas accents are almost never marked by short duration. An integral part of such a pitch configuration is thus either a 'before–after' relation in a phonological model, or a parametric representation of the position of prominent points on the time axis; therefore, we will call pitch *bi-dimensional*. These positions as such do not, however, represent pitch but duration — simply because they are measured in msec, and because they are often highly correlated with duration features [1]. We will therefore distinguish between 'genuine' pitch features (F0) and durational features (DUR) which model pitch position or trajectories on the time axis. Our DUR features do not represent fully the parameter 'duration' but only the temporal aspects of pitch configurations.

5. FEATURE EXTRACTION AND CLASSIFICATION

We computed 281 features (238 modelling F0 and 43 modelling DUR) using two different basic approaches: in the first, values were computed directly over the whole chunk, such as slopes and regression coefficient with mean square error, functionals covering the first four central moments as well as extreme values such as maxima or minima, range, and positions on the time axis, ratio of voiced and unvoiced speech, magnitudes and steepness between adjacent extrema. In the second, word-based and sequential approach, functionals such as mean, standard deviation, minimum, maximum, linear regression coefficients and errors as well as positions on the time axis were extracted for each word; then, minimum, maximum, and mean functionals were computed for each chunk.

Table 2: Classifi cation results in F (F0, DUR, and F0+DUR) for a 2- and a 4-class problem, computed with SVM and RF: *aut*omatic extraction vs. manual *corr*ection

feature types	SVM		RF	
	aut	*corr*	*aut*	*corr*
prominence: two classes				
F0	73.9	75.7	74.8	75.5
DUR	74.5	74.2	75.2	75.2
F0+DUR	76.5	78.1	76.4	78.2
emotion: four classes				
F0	53.2	54.5	54.7	55.3
DUR	50.9	51.1	52.3	52.5
F0+DUR	54.5	55.8	57.3	57.5

The data was partitioned into three balanced splits meeting the following requirements (in order of priority): no splitting of within-subject chunks, similar distribution of labels, balance between the two schools, and balance between genders. In order to have a more balanced distribution for training, we upsampled all classes but **N***eutral*: 3x **M***otherese*, 2x **E***mphatic*, and 2x **A***ngry*. We computed a 3-fold cross-validation with support-vector-machines SVM and Random Forests RF. Results are given in Table 2 where we report the F value which is used in the interest of having a unique performance measure; here, F is defined as the uniformly weighted harmonic mean of RR and CL: $2 \cdot CL \cdot RR/(CL+RR)$. RR is the overall recognition rate (number of correctly classified cases divided by total number of cases or weighted average); CL is the 'class-wise' computed recognition rate, i.e. the mean along the diagonal of the confusion matrix in percent, or unweighted average. The F measure represents a trade-off between CL and RR.

It could be expected that (1) overall performance and trends are similar across the two classifiers; higher performance can be expected if more than only the two types of parameters F0 and DUR are used [3]. (2) If F0 is involved (F0, F0+DUR),

corr always yields better results than *aut*; the differences are, however, not that big as one might expect. Note that for both *corr* and *aut*, manually corrected word segmentation has been used. (3) In relation to F0, DUR is more important for the 2-class problem **prominence** than for the full 4-class problem **emotion**. This is in accordance with [1, 5]; it might as well be due to specifities of the four emotion classes. (4) The combination of F0 and DUR in F0+DUR contributes to performance; we can speculate whether this is due to a better modelling of pitch configurations, or 'simply' due to the fact than information pertaining to two different types of parameters is being used. (5) We should keep in mind that DUR does not fully model duration, and that there are much less DUR than F0 features; a full modelling of duration might result in higher relevance.

The caveat has to be made that we are not talking about very pronounced differences; the experiments should thus be repeated with other data and other features. Due to the problem of repeated measurements [3], we refrain from interpreting differences in terms of significance.

To assess tentatively the importance of features, we computed a Prinicipal Component Analysis (PCA) for the 4-class problem, and used the PCs as features in classification (SVM with Sequential Forward Floating Selection); for each of the top three PCs for each split, we had a look at the three most important features. For F0+DUR, mean F0 is more important for *aut* than for *corr*; DUR (here, position of F0 offset in the words) is more important for *corr*. Especially for *corr*, features modelling the slope of the pitch contour (such as kurtosis and regression) seem to be most important. This difference might illustrate the drawbacks of automatic F0 extraction which produces errors that 'smear' word-final values and characteristic contours; thus for *aut*, the more robust mean values come to the fore instead.

6. DISCUSSION AND CONCLUDING REMARKS

We have shown that at least for our data and classifiers, and using the ESPS algorithm, F0 extraction errors affect classification performance, albeit not to a very high degree. If this can be corroborated with other data and other algorithms, it is reassuring — it would mean that not much gets lost if we extract pitch automatically. Of course this does not yet give the final answer to the fundamental question whether the (possibly) low relevance of pitch is a 'bug' (due to erroneous extraction) or a feature (due to pitch simply being less relevant, compared to the other parameters): it can be both. If it is a 'bug' it might be difficult to overcome: there is no error-free F0 extraction algorithm, and it is normally too much effort to correct F0 values manually when dealing with large databases.

We have also illustrated the intertwining of prosodic parameters – here, of pitch and duration. Pitch values alone do of course not completely represent pitch as a phenomenon because the constellation/tones/contours that constitute pitch necessarily entail some temporal information. The other parameter, duration, so to speak 'steals in' which means in turn, that a full modelling of pitch as a complex phenomenon (i.e. pitch contours) is represented not only by pitch (F0) but by duration information as well. This of course is no problem if we simply accept a certain inseparability of speech parameters.

7. ACKNOWLEDGMENTS

The initiative to co-operate was taken within the European Network of Excellence HUMAINE under the name CEICES (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates). 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 responsibilty lies with the authors.

8. REFERENCES

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

[2] A. Batliner, S. Burger, B. Johne, and A. Kießling. MÜSLI: A Classifi cation 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, 1993.

[3] 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 Classifi cation of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana, 2006.

[4] O. Dioubina. Annotation of expressive speech. In *Proc. VOQUAL'03*, pages 173–177, Geneva, 2003.

[5] 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.

[6] S. Mozziconacci. *Speech variability and emotion: production and perception*. PhD thesis, Technical University Eindhoven, 1998.

[7] 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.