In “Speech Recognition and Coding — New Advances and Trends”, Rubio–Ayuso and Lopez–Soler (eds.), Springer, 1995, pp. 101–104.

1

Can You Tell Apart Spontaneous and Read Speech if You just Look at Prosody?

A. Batliner[1]
R. Kompe, A. Kießling, E. Nöth, H. Niemann[2]

ABSTRACT Although the recognition of spontaneous speech is the ultimate aim of speech understanding systems it has rarely been investigated so far. In this article first analyses of a German database containing identical utterances of spontaneous and read speech are presented. We describe the differences in prosody between these two registers and report results of a classifier that was trained using prosodic features to discriminate spontaneous and read speech. A systematic difference could be observed that is however rather complex and partly speaker dependent.

1.1 Introduction

Up to now, most research on automatic speech understanding (ASU) in general and on the use of prosodic information for this task in particular has been done on read speech (i.e., non-spontaneous speech, henceforth NSP) from experienced speakers. The ultimate aim, however, of ASU is the recognition of spontaneous speech (henceforth SP) from naive speakers. Nevertheless, the advantage of elicited NSP is obvious: it is possible to obtain easily huge databases that contain exactly the phenomena one is interested in. In contrast, it is very time consuming to obtain and transliterate SP. The question now is whether one can use NSP for training and SP for recognition. A necessary prerequisite is the investigation of systematic differences between SP and NSP. As we are interested in the use of prosodic information in our ASU system, we will therefore investigate which differences can be found between SP and NSP concerning prosodic features. Results are obtained using three strategies. One is simply to compare the mean values of the features. Second the correlation between the features and a spontaneity judgment of listeners is computed. Third a classifier is trained automatically using these features to discriminate SP and NSP. If the classifier performs reasonably well, one can conclude that the features differ between SP and NSP.

1.2 Material and experimental design

First analyses of a German database containing utterances of SP and parallelized NSP are presented. Two pairs of speakers (3 female: C, X, and A, 1 male: F) had to solve problems in a “blocks world”. The experiment was designed in a way that resulted in absolutely SP (short clarification dialogs with many turn takings). Those utterances were chosen that had a

[1]L.M.-Univ. München, Institut für Deutsche Philologie, 80799 München, F.R. of Germany.
[2]Univ. Erlangen-Nürnberg, Lehrstuhl f. Mustererkennung (Inf. 5), 91058 Erlangen, F.R. of Germany
[3]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.

sufficient signal quality and did not contain specific non-syntactic phenomena like hesitations which are normally only found in SP (otherwise listeners in the perception experiments could have been guided by these phenomena to differentiate between the registers, cf. below). After 9 months, the same speakers read the chosen utterances, their own and those of the partner, given in written form and embedded in a sufficiently large context. The written utterances were not given in the orthographically correct form but in colloquial speech, thus approximating SP. The parallelized SP and NSP utterances are therefore as similar as possible with respect to segmental information. Our experimental design guarantees that the NSP utterances are produced as "spontaneously" as possible and is thus "conservative" in the sense that the object of investigation – the difference in register – is difficult to produce. In this special NSP register, the usual transformation of the canonical, orthographic form into speech is thus missing but of course none of the other planning processes that characterize reading. Our material can be taken as representing two prototypical registers. Note, however, that in speech there exist not only two registers but many different, some of them being more, others less spontaneous. Since in dialogs humans are able to adopt to the speaking register of the partner, this ability might also be useful in ASU in the future. In this paper, we will not deal with the utterances that were read by the partner. Informal listening tests showed that three of the speakers read 'very spontaneously' whereas the fourth one (A) had a clear shift in register. Recording conditions were comparable to a quiet office environment. The utterances were digitized with 12 Bit and 10 kHz; a total of 886 utterances (without the reread partner utterances, i.e. C: $2\star181$, X: $2\star118$, A: $2\star66$, F: $2\star78$) that amount to about 18 minutes of speech material. For more details cf. [1].

1.3 Perception experiments and extracted features

In perception experiments, 10 subjects judged the degree of spontaneity of each utterance on a scale from 1 ("very spontaneous") to 4 ("not spontaneous at all"). NSP and SP utterances were presented in random order; only the utterances of one speaker were presented per session. For the evaluations described below the average of the judgment of the 10 subjects was used. Using three different $F0$ algorithms, $F0$ was computed and corrected manually to obtain a reference contour. A time alignment with the canonical pronunciation and a broad transcription was done with an automatic speech recognizer and corrected manually as well. From the corrected $F0$ contour, which is given in semi-tones to the basis 1 Hz, the following features were extracted: onset, offset, maximum, minimum, range, mean, standard deviation, and regression coefficient which is given in semi-tones per second. These features were normalized with respect to the utterance specific pitch level (subtraction of the mean value – of course, the regression coefficient and the mean value itself are not normalized). For each utterance a measure for the speaking rate was computed: the average of the actual phone duration divided by the intrinsic duration (average duration of the same phone in the whole database) with the formula given in [5].

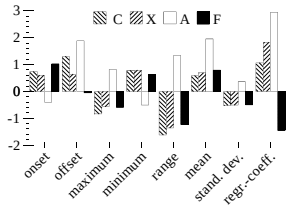

FIGURE 1.1: Differences between the averages for SP and NSP for the $F0$ features (subtraction of NSP from SP values)

1.4 Results and discussion

In figure 1.1 and figure 1.2 (left), differences of the averages of the $F0$ features of SP and NSP and of the speaking rate respectively are given for each speaker separately. NSP values are subtracted from SP values. In figure 1.2 (left) the value of -0.11 for speaker X e.g. means that on the average the phones in SP are 11% shorter than in NSP. Globally, the difference between the two registers can be characterized as follows. In SP, onset, offset, and minimum are higher and the maximum is lower than in NSP. The regression coefficient and the mean are higher, the range and the standard deviation are lower in SP. SP is faster than NSP. Obviously these global differences hold rather systematically for C and X; F and especially A sometimes behave differently. This different behavior can also be expressed by the following rating: For each feature in figures 1.1 and 1.2 (left) the mean of the 4 values was computed. Then the speakers were ranked, i.e. for each feature we assign a number from 1 to 4 to each speaker, where 1 means that the value corresponding to the speaker is closest to the mean. Over all features speaker A got an average rank of 3.4 (F: 2, C: 1.9, X: 1.6), which is in good agreement with figure 1.2 (right) where the difference between the averages of the spontaneity judgment of NSP and SP is shown. According to the listeners for all speakers the degree of spontaneity is higher for SP than for NSP utterances. Yet the differences are speaker dependent and greatest for speaker A who has a real shift in register according to the informal tests. There is a dependency between the judgment and the prosodic features, but not a very strong one, R^2 ("percentage of variance explained") being between 0.21 and 0.40.

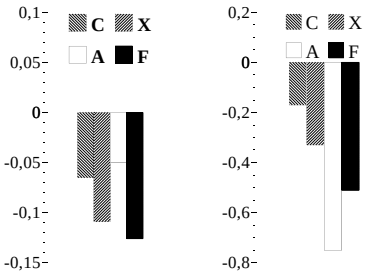

FIGURE 1.2: left: Difference between the average of the speaking rates on SP and NSP; right: Difference between the average of the spontaneity judgment on SP and NSP

For the classification of SP vs. NSP, discriminant analyses were conducted. In figure 1.3 classification results for multi-speaker (all) and speaker-dependent (C, X, A, F) experiments using the same training and testing data are given. In one set of experiments all prosodic features (black bars) were used for classification, in the other experiments the average spontaneity judgment was used (white bars). In the speaker-dependent experiments the recognition rate based on the prosodic features is between 74% and 79%. In multi-speaker mode the rate goes down to 65%. The recognition rates based on the prosodic features are much higher for C, X and F than those based on the spontaneity judgment. Only for speaker A is the spontaneity judgment slightly better. The recognition rates based on the spontaneity judgment conform with the differences between the judgment on SP and NSP as shown in figure 1.2 (right). With learn $\neq$ test (2/3 of the utterances of one speaker for training and the other 1/3 for testing), the results were about 3% worse. If only a single prosodic feature was used for classification the recognition rates were much lower (between

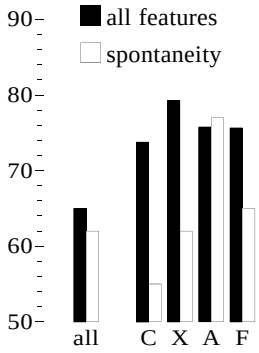

FIGURE 1.3: Results of the classification in SP and NSP based on the prosodic features

55% and 65% in speaker-dependent mode); this holds for all features. Our results show that the difference between SP and NSP is rather complex. The averages of a single feature are mostly not markedly different between SP and NSP (cf. figure 1.1 and 1.2 (left)), but when taking all these features together a classifier is able to distinguish quite well between SP and NSP. Thus there seems to be a difference in the overall prosody between SP and NSP, indicating the shift in register. Yet this shift seems to be different from the shift in register the subjects of the tests perceived, because it correlates much more with the prosodic features than with the spontaneity judgment. Note, however, that the listeners had to judge the degree of spontaneity of each utterance and not to assign it to either SP or NSP.
There are a few other comparable studies, whose results differ partly from our results; for details cf. [1]. In [3] e.g. it is reported that in American English $F0$ range is **greater** in SP than in NSP. For the moment it cannot be decided whether language-, speaker-, design-, or register-specific factors are responsible for these differences.

1.5 Final remarks

We have shown that one can tell apart SP and NSP reasonably well if one just looks at prosody. However, this difference has to be investigated further, because it is not trivial. In particular, there are not only SP and NSP but many different registers influenced by many parameters such as speaker, speaker-partner relationship, read or non-read, dialect, etc. If one wants to use prosodic features in ASU it might be difficult to train the system parameters on NSP and to test on SP as it is often done. On the other hand, it is shown in [2, 4] that e.g. the prosodic marking of questions vs. non-questions is generally more distinct in SP but not categorically different from NSP. A classifier that was trained on NSP showed good results for SP. The answer to the question whether one can use NSP for training and SP for testing might be a clear “yes, but – be careful”.

1.6 References

[1] A. Batliner, B. Johne, A. Kießling, and E. Nöth. Zur prosodischen Kennzeichnung von spontaner und gelesener Sprache. In G. Görz, editor, *KONVENS 92*, Informatik aktuell, pages 29–38. Springer-Verlag, Berlin, 1992.

[2] 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 *Proc. ESCA Workshop on prosody*, pages 112–115, Lund, September 1993.

[3] N. Daly and V. Zue. Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech. In *Int. Conf. on Spoken Language Proc.*, vol 1, pages 763–766, Banff, 1992.

[4] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. *“Roger”, “Sorry”, “I’m still listening”: Dialog guiding signals in information retrieval dialogs* . In *Proc. ESCA Workshop on prosody*, pages 140–143, Lund, September 1993.

[5] C. Wightman and M. Ostendorf. Automatic Recognition of Intonational Features. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages I–221–I–224, San Francisco, 1992.