Emotional Aspects of Intrinsic Speech Variabilities in Automatic Speech Recognition

Miloš Cerňak and Christian Wellekens

Institut Eurecom, Sophia-Antipolis, France
Milos.Cernak@eurecom.fr and Christian.Wellekens@eurecom.fr

Abstract

We analyze two German databases: the OLLO database [1] designed for doing speech recognition experiments on speech variabilities, and the Berlin emotional database [2] designed for the analysis and synthesis of emotional speech. The paper tries to find a relation between intrinsic speech variabilities and the emotions. Moreover, we study this relation from the point of view of speech recognition. Acoustical analysis is performed on both databases, using Normalized Amplitude Quotient and F0 parameterization of five analyzed vowels [a], [e], [i], [o], and [u], merging their long and short variants. Euclidean distance between the feature vectors of both databases is used for finding the relation, named as emotional aspect of speech variabilities. The speech recognition experiments on the OLLO database show that found emotional aspects have also a discrimination power.

1. Introduction

Research effort on emotions and speech processing falls into a field what is called 'affective computing' - the goal is to design ASR and TTS related algorithms that understand and respond to human emotions [3]. Usually a paralinguistic approach is used for the analysis of emotional state of the speaker. It uses different components of the raw acoustic signal to infer underlying emotion states of the speaker. It is needed first to carry out basic signal processing operations, next to use their outputs to describe the main prosodic structures present in a general way, and then to use those descriptions to find parameters that may be relevant to specific emotions.

Within the DIVINES project, we mainly study deficiencies of speech recognition process due to speech intrinsic variabilities. However, all these variabilities were till now not studied from the emotional point of view. We have found interesting to know how emotions encoded in speech influence the accuracy of computer speech recognizer. However, a major obstacle to developing accurate models of human emotion is still the absence of rich, realistic emotional databases, and moreover, which should be suitable for performing speech recognition experiments. Indeed, confirmed by our preliminary work on that topic, the Berlin emotional database [2] (henceforth 'EMO' database) cannot itself be used for speech recognition experiments due to lack of enough data.

While in speech recognition most of the effort was devoted to classification of a speech signal into its correct emotion category, less attention was devoted on recognition of speech produced in different emotional states. This paper would like to contribute to the second mentioned area. The aim of this study is to find a relation between intrinsic speech variabilities and the emotions. Two German databases are analyzed for that purpose. The OLLO database [1] is indexed by x_i (see Eq. 4) in terms of six speaker-dependent speech variabilities, including **fast** and **slow** speaking rate, **statement** and **questioning** styles, and **high** and **low** speaking effort. The EMO database [2] is indexed by y_j (see Eq. 4) in terms of six emotions, including four major emotions **anger**, **happiness**, **fear** and **sadness**, plus **boredom** and **disgust**. The acoustical analysis yields the prosody parameterizations, and the Euclidean distance is employed to find a link (relation) between the intrinsic speech viariabilities of the OLLO database and the emotions of the EMO database.

The paper is structured as follow. Next Section 2 describes analyzed material of both databases. The overview of acoustical analysis is given in Section 3. Section 4 introduces a technique to assign emotional aspects to speech variabilities. Section 5 describes speech recognition experiments done on the OLLO database from the point of view of emotional aspects, and finally Section 6 concludes the paper and presents possible future work.

2. Analyzed material

This section describes used speech databases. Both databases have different phonesets. The phoneset of EMO database consists of more than 40 phonemes, while the phoneset of OLLO consists of 25 phonemes. Because the utterances of the OLLO databases does not have supersegmental structure, we were forced to do our analysis on **segmental level**. For our joint analysis of both databases we have found useful to do the analysis of the short and long vowels, which were sufficiently represented in both databases.

The OLLO database is designed for recognition of in-

dividual phonemes that are embedded in logatomes, specifically, CVC and VCV sequences. Several intrinsic variabilities in speech are represented in OLLO, by recording from 40 speakers from four German dialect regions, and by covering three speaker-dependent variabilities: gender, age and dialect, and six speaker-dependent variabilities. The analyzed material of the OLLO database, used in this work, consists of almost 27000 logatomes, a list designed for speaker independent speech recognition experiments as provided by release version 3.1 of the OLLO. Almost 39500 vowel realizations were selected for the analysis.

The EMO database is designed for doing analysis of emotional speech. It consists of ten utterances that were produced by five male and five female German actors enacting the emotions and neutral version. The sentences were taken from everyday communication and could be interpreted in all emotional contexts without semantic inconsistency. Kienast [4] presents more detailed study of the vocal expression of these emotions, showing various specific characteristics concerning vowel quality, segmental reduction and energy distribution in voiceless fricatives. The database was evaluated by Burkhardt [2] in an automated listening test and each utterance was judged by 20 listeners with respect to recognisability and naturalness of the displayed emotion. For our analysis we used a subset of 493 utterances, recognized by listening tests better than 80% and judged as natural by more than 60% of the listeners. In total 5100 vowel realizations were selected for the analysis.

3. Acoustical analysis

As the role of emotions in speech can be viewed as an emotional prosody (see e.g. [5]), we analyzed the prosody of 5 vowels: [a], [e], [i], [o], [u], merging their long and short realizations into a single category. A description of emotional speech unavoidably runs into the problem of distinguishing between prosodic and paralinguistic features of speech [6]. Campbell [7] has recently showed that voice quality or tone-of-voice is controlled in much the same way as the more traditional prosodic parameters of intonation, amplitude, duration, and timing. For our analysis we used pitch range that belongs to the prosodic features, and voice quality that belongs to the paralinguistic features. We used Normalized Amplitude Quotient (henceforth 'NAQ') features for voice quality representation, as it has been found to be very robust parameterization of the glottal flow [8].

The amplitude quotient AQ is computed as:

$$\mathrm{AQ} = f_{ac}/d_{peak}, \tag{1}$$

where f_{ac} denotes ac flow, and d_{peak} denotes negative peak amplitude of the differential flow. The AQ quantifies the closing phase of the glottal flow, and it reflects changes that occur in the glottal source when vocal intensity or phonation type is altered [8]. The time-length measure given in Eq. 1 can be normalized with respect to the length of the pitch period:

$$\mathrm{NAQ} = \mathrm{AQ}/T, \tag{2}$$

where T denotes the length of a pitch period.

The glottal volume velocity waveform was estimated by voice inverse filtering [9]. Static parameters were used for that estimation: 31 model order for vocal tract, and discrete all-pole modeling for AR-modeling technique. Then, NAQ features were calculated using the HUT Aparat toolbox [10].

The statistical tests were done on the acquired emotional prosody parameterization. Two-way ANOVA tests with one independent variable for the vowel categorization and one independent variable for the emotions categorization confirmed highly significant differences. The NAQ and F0 features were significantly different at a 99 % confidence level (except the F0 features for men in the EMO database, where differences between the vowels were significant at a 95 % confidence level).

4. Finding the relation

For each vowel in the databases we calculated one NAQ value and one F0 value per emotion and one NAQ value and one F0 value per speech variation. We gathered all the values per vowel in the databases and normalized them using z-score:

$$Z = (X - M)/S, \tag{3}$$

where Z refers to the z-score, M is the parameterization's mean, S is the parameterization's standard deviation, and X is an individual score of NAQ and F0 within the distribution having mean M and variance S. The value of z-score is difference from mean in SD units.

We further used medium value of Z as a vowel representative value. It allows us to create two 5-dimensional feature vectors composed of representative values for each speech variability and each emotion. The first feature vector is composed of NAQ values, and the second feature vector is composed of F0 values. Then, we assigned an emotion to each variation of the OLLO database. Some authors use a Mahanalobis distance [11] for that purpose. Because the dynamic range of the feature set was already normalized using z-score, we used simple Euclidean for finding the relation in the following way:

$$d(x_i, y_j) = d(x_i^{\mathrm{NAQ}}, y_j^{\mathrm{NAQ}}) + d(x_i^{F0}, y_j^{F0}), \tag{4}$$

where $x_i = (x_i^{NAQ}, x_i^{F0})$ denotes the index to the OLLO database in terms of the i-th speech variation, and $y_j = y_j^{NAQ}, y_j^{F0})$ denotes the index to the EMO database in terms of the j-th emotion. The lowest distance $D_{i,j}$ was

Speech variability	Emotional aspect
Fast, loud, quiet	Happiness
Slow	Sadness
Questioning style	Anxiety/fear
Statement style	–

Table 1: Link of the emotional aspects and the speech variabilities for women.

Speech variability	Emotional aspect
Fast, loud, quiet	Boredom
Slow	Neutral
Questioning style	Boredom
Statement style	Neutral

Table 2: Link of the emotional aspects and the speech variabilities for men.

then used to assign j-th emotional aspect to i-th speech variation:

$$D_{i,j} = \arg\min_{j} \left(d\left(x_i, y_j\right) \right). \quad (5)$$

Tables 1 and 2 show relation between the emotional aspects and the speech variabilities, evaluated separately for the women and men.

Different emotional aspects were assigned to women and men. It is a commonly held belief that men and women treat their emotions in different ways (see e.g. the conclusion of the Tickle's study [12]). Our study confirmed that the emotions of happiness, sadness and fear are more characteristic of women. Research also suggests that men express less anxious and depressed feelings than women do.

We have not assigned the emotional aspect to the 'normal' speech variability for the women, while no one of computed distances to the emotional feature vectors was reasonably low (or better said, comparably low as the other minimal distances for the rest of speech variabilities).

5. Experiments with automatic speech recognition

Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition system were trained using public domain speech library TORCH [13] on the NO-accent training part of the OLLO training set that consist of 13446 logatome utterances[1]. Five states left-right HMM models were trained for each of the 26 phonemes in the OLLO database including silence as well. Gaussian mixture models with 17 Gaussians per state and diagonal covariance matrices were used to model the emission probability densities of the 39 dimensional feature vectors (13 MFCC + 13 deltas + 13 accelerations). Phoneme recognition rates reported over each variability are given in Fig 1.

Comparing recognition accuracies from the point of view of the emotional aspects, see table 3 (happiness for females and boredom for males), we can see that the emotional aspects have also a discrimination power. Splitting the recognition results according to the emotional aspects (italicized lines versus non italicized lines in the table 3) yields significant different accuracies at a 95% confidence level (paired t-test, $p = 0.00232$). Slow speech and speech with questioning or statement style show significantly better accuracy. On the contrary, we may conclude that we can not discriminate different accuracies within the same emotional aspect, as it is for fast, loud and quiet speech.

Questionable is the emotional aspect 'Boredom' for questioning style of men. We tried to find an intersection (see table 3) between speech variabilities and women and men's emotional aspects. The assignment of 'Boredom' for questioning style of men was therefore excluded from the intersection. We backtracked this assignment, finding that calculating distance only with NAQ features, the assignment would be 'Anxiety', the same as for women. However, adding the distance of F0 features resulted in the emotional aspect 'Boredom'. The problem is maybe in F0 features, which are quite different for women and men, and maybe some compensation of that should be applied. For now we don't have a solution for that.

6. Conclusion

The novel contribution of this paper is in finding relation between intrinsic speech variabilities and emotions, naming this relation as emotional aspect of intrinsic speech variability. Discrimination power of emotional aspect of the speech variability was shown in predicting significantly different recognition accuracy. We imply that the emotional aspects might be useful indicators of recognition failures.

Prosodic features have also been shown to be effective in ranking recognition hypotheses, as a post-processing filter to score ASR hypotheses [14],[15]. In the future we would like to study if both prosodic features, NAQ and F0, might be used together with an acoustic-based confidence score, to improve a decision to reject a recognition hypothesis.

7. Acknowledgment

This work has been supported by EC 6^{th} Framework project DIVINES under the contract number FP6-002034.

[1]We did several comparative studies of the performance of the speech recognizer based on TORCH with public domain speech recognizer HTK, which resulted in comparable results for that domain

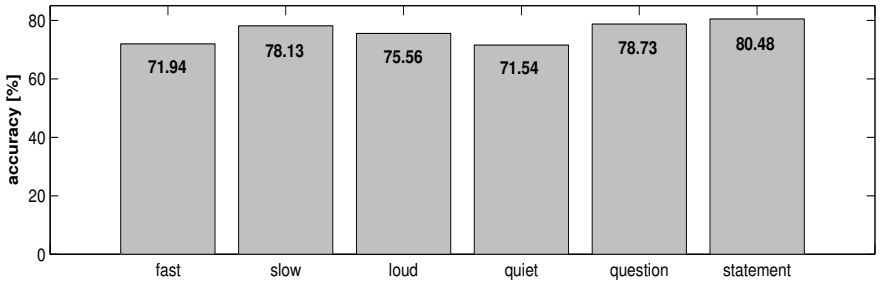

Figure 1: Phoneme recognition rates reported over each variabilites

Speech variability	Recognition rate	Emotional aspect - W	Emotional aspect - M
Fast	*71.94*	*Happiness*	*Boredom*
Slow	78.13	Sadness	Neutral
Loud	*75.56*	*Happiness*	*Boredom*
Quiet	*71.54*	*Happiness*	*Boredom*
Questioning style	78.73	Anxiety/fear	Boredom
Statement style	80.48	–	Neutral

Table 3: Recognition rates from the point of view of the emotional aspects.

8. References

[1] Wesker, T., Meyer, B., Wagener, K., Anemuller, J., Mertins, A., Kollmeier, B.: Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. In: Interspeech 2005. (2005) 1273–1276

[2] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Interspeech 2005. (2005) 1517–1520

[3] Bosch, L.T.: Emotions: What is possible in the asr framework. In: ITRW on Speech and Emotion. (2000) 189–194

[4] Kienast, M., Sendlmeier, W.F.: Acoustical analysis of spectral and temporal changes in emotional speech. In: ITRW on Speech and Emotion. (2000) 92–97

[5] Yang, L.C.: Prosodic shape and expressive meaning: The expression and recognition of emotions in spontaneous speech. In: Prosody in Speech Recognition and Understanding. (2001)

[6] Roach, P.: Techniques for the phonetic description of emotional speech. In: ITRW on Speech and Emotion. (2000) 53–59

[7] Campbell, N.: Accounting for voice-quality variation. In: SP-2004. (2004) 217–220

[8] Alku, P., Backstrom, T.: Normalized aplitude quotioent for parametrization of the glottal flow. Journal of the Acoustical Society of the America **112**(2) (2002) 701–711

[9] Alku, P.: Parameterisation methods of the glottal flow estimated by inverse filtering. In: Voice Quality: Functions, Analysis and Synthesis. (2003) 81–88

[10] Airas, M., Pulakka, H., Backstrom, T., Alku, P.: A toolkit for voice inverse filtering and parametrisation. In: Interspeech 2005. (2005) 2145–2148

[11] Amir, N., Ron, S., Laor, N.: Analysis of an emotional speech corpus in hebrew based on objective criteria. In: ITRW on Speech and Emotion. (2000) 29–33

[12] Tickle, A.: English and japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In: ITRW on Speech and Emotion. (2000) 104–109

[13] Collobert, R., Bengio, S., Marithoz, J.: Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP (2002)

[14] Hirose, K. In: Disambiguating recognition results by prosodic features. Springer (1997) 327–342

[15] Hirschberg, J., Litman, D., Swerts, M.: Prosodic and other cues to speech recognition failures. Speech Communication **43**(1-2) (2004) 155–175