CROSS-CULTURAL RECOGNITION OF AUDITIVE FEEDBACK USING ECHO STATE NETWORKS

Anja K. Philippsen, Kai A. Mismahl, Britta Wrede and Yukie Nagai

Bielefeld University (Germany), Osaka University (Japan)
aphilipp@techfak.uni-bielefeld.de

Abstract: This paper deals with the development of a classifier to distinguish between positive and negative feedback from the user in human-machine-interaction. We calculate prosodic features from the user's utterances and feed it to an Echo State Network, a dynamic classifier that is able to learn temporal dependencies implicitly. The data were recorded in a test scenario from German and Japanese test subjects, once in natural speech and once in an artificial "language" that uses only the syllable "na". The test subjects had to give feedback to a simulation of the robot Flobi and were instructed to behave like interacting with a child. The implemented Echo State Network proved to be able to learn to classify the feedback of a single person into the two categories "positive" and "negative" and could generalize to a certain extent. We experience a high range of different feedback in the data, intra-culturally as well as inter-culturally. However, it can be shown that a classifier trained on German data works significantly better on German data than on the Japanese, indicating that cultural differences exist. Analyzing different feature subsets, we found out that using Mel-Frequency Cepstral Coefficients as features yield a better classification rate than using prosodic features (like pitch and intensity) alone.

1 Introduction

Feedback from the user is an important information for a successful interaction, in human-human as well as in human-machine-interaction. In the same way children learn through the feedback of their caregivers, it would be desirable that e.g. robots could use feedback information to adjust their behavior.

While the content of an utterance is defined by the lexical information, the prosody is supposed to convey the emotional state of the interaction partner. This additional information is usually not exploited by speech recognition systems, although lexical information alone are not always complete. E.g. in sarcastic or in informal speech questions might not be marked by grammatical constructs but only through intonation. Prosody is especially helpful if lexical information is not available, e.g. in a foreign language.

A study from Fernald [4] showed that preverbal infants can distinguish between approval and prohibition using prosodic information in infant-directed speech. In contrast to adult-directed speech infant-directed speech shows differences in fundamental frequency, intensity and duration of the utterance. This phenomenon can be found across several languages, including German and Japanese [5]. In the above study [4] the English infants could classify infant-directed speech in several unfamiliar languages, including German, but were not able to classify infant-directed Japanese speech.

While preverbal infants rely on prosody alone, they tend to prefer lexical content as their vocabulary grows [6]. Around the age of ten children start to prefer prosodic contents again, if

lexical and prosodic contents are discrepant [7]. When facing utterances in a foreign language, children of all age are able to use prosody to classify the included emotion [11].

Using data recorded from German as well as from Japanese test subjects we aim at a cross-cultural analysis.

In [2], Breazeal and Aryananda differentiated between four categories of prosodic intention in robot-directed speech (their robot was earlier shown to evoke infant-directed speech): praise, prohibition, attention and soothing. They were able to recognize utterances of these categories using the pitch contour and some static features of the signal like mean pitch, intensity and duration. Their algorithm works decently well on the caregivers' utterances as well as on utterances selected because of the affective strength.

In our research we distinguish two categories, deciding whether the human is content with the robot's behavior (positive feedback) or not (negative feedback). These are important for the robot to decide whether his action was correct. A qualitative analysis of our data indicated that positive feedback often matches with the categories praise and attention (higher mean pitch and higher pitch change), while the negative feedback has lower mean pitch and less pitch change like prohibition and comfort. But there existed high variance; the results were not consistent with all test subjects (note, that we did not carry out a preselection like in [2]). As a classifier we, therefore, used an Echo State Network (ESN), a type of a Recurrent Neural Network. This dynamic classifier has similar properties like Hidden Markov Models (HMMs) and has also been shown to be able to take the role of HMMs in speech recognition tasks [12].

Our research questions are the following:

1. Is it possible to build a classifier to distinguish positive and negative feedback in a speaker-dependent / speaker-independent way?

2. Is there a difference in how Japanese and German people use prosody to convey feedback?

2 Data Recording

Feedback could serve as an important clue for robots to learn the correct names of objects. We used a robot simulation of Flobi [10] as an interaction partner. Flobi was naming objects presented as an image on a computer screen and the human should respond to Flobi whether he was correct or wrong. The scenario is shown in Figure 1. The robot's utterances have been recorded from an Asian woman in German and from a German woman in Japanese. By using non-native speakers the scenario should appear more plausible, as if the robot learned words from a foreign language. The test subjects were instructed to behave as if they were correcting a child. While some test subjects used clearly motherese, others answered in a less affective way.

Figure 1: A Japanese test subject gives feedback to Flobi during data recording.

Overall we had 32 test subjects, 22 German (11 female, 11 male) and 10 Japanese (2 female, 8 male). For each person we carried out two runs containing 30 objects each. While in one of the runs they were allowed to say whatever they wanted (even multiple sentences), in the second

run they were limited to use one syllable (“na”) and only change the prosody to convey the positive or negative meaning. The idea was to force the usage of prosody through the absence of lexical information. Another advantage of this method is that we get data that are independent of linguistic characteristics of a specific language.
The data were recorded with an headset at a sample rate of 48000 Hz.

3 Methods

In the following we describe the methods we used for speech extraction and classification. All values mentioned were calculated on windows of 4096 samples of the signal (i.e. approx. 85ms). Subsequent windows have an overlap of 75%.

3.1 Speech Extraction

To extract the training data, i.e. to detect speech in the signal, we used in addition to the pitch value another value, which is in literature often called the *periodicity* [13]. This value can be interpreted as the proportion of harmonics in the signal; the more distinct the maximum of the cepstrum of the speech snippet is, the higher is the periodicity value and the more likely a voiced speech segment is present.
Looking for a high periodicity and a pitch value in a plausible range (50 Hz to 500 Hz for human speech) we can cut out speech automatically from the signal.
There still might be gaps in the signal, e.g. because consonants have a low periodicity, so we used opening and closing algorithms (usually used in image processing to get rid of single noise pixels) to throw away short as speech classified snippets inside noise or as noise classified snippets inside a speech segment.
As there is no clear threshold, which periodicity value has to be reached, we conducted an Expectation Maximization Algorithm to enhance the results: The features of the signal are plotted in a 2-dimensional (3-dimensional) feature space (using maximum and variance of the cepstrum (and logarithmic energy) as features). In this space two clusters are searched (speech and non-speech) and each sample in the signal is relabeled according to which cluster it belongs to. Then again outliers in the signal are removed using the opening and closing algorithms.
In the following analysis we only included utterances extracted by this algorithm; false positives (extracted segments that do not contain speech) were removed manually.

3.2 Feature Extraction

Prosody is usually defined as a variation of the fundamental frequency (pitch), the intensity and the duration of an utterance [1]. We used a dynamic classifier, which is able to model durational characteristics, so that we do not need to include them explicitly.
Our feature vector consists of the log energy and pitch of the signal, the mean, variance and maximum amplitude of the cepstrum, and 12 mel-frequency cepstral coefficients, as well as the first and second derivative of all features.

3.3 Classification Using Echo State Networks

ESNs have been proposed by Herbert Jaeger in [9]. In contrast to general recurrent neural networks all connections, except those to the output neurons, are chosen randomly and fixed. This so called reservoir creates non-linear random projections of the input vector and serves as a short term memory. There is no need to propagate the error back through time while learning, which reduces the computational effort.

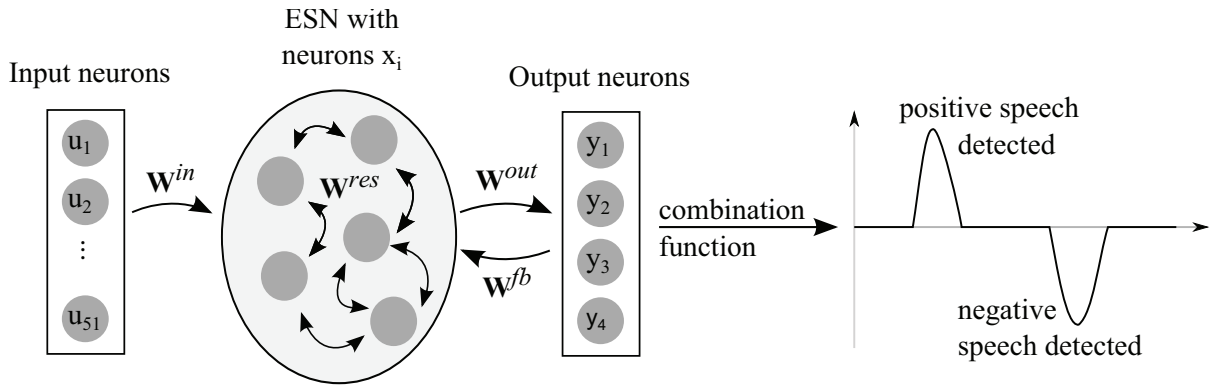

Figure 2: The information flow of our classifier: The input vector is consecutively fed into the reservoir, the internal state and then the readout is calculated. The output neurons are fed back into the network, and are also combined to produce a single output.

As depicted in Figure 2 the next state $\mathbf{x}_{n+1}$ of the reservoir is calculated as a sum of the next input $\mathbf{u}_{n+1}$, the current reservoir state $\mathbf{x}_n$ and the current output vector $\mathbf{y}_n$, all multiplied with the according weight matrices [9]:

$$\mathbf{x}_{n+1} = f(\mathbf{W}^{in} \cdot \mathbf{u}_{n+1} + \mathbf{W}^{res} \cdot \mathbf{x}_n + \mathbf{W}^{fb} \cdot \mathbf{y}_n)$$

For f, the activation function, we used the hyperbolic tangent. Note, that the weight matrices $\mathbf{W}^{in}$ for the input, $\mathbf{W}^{res}$ for the reservoir and $\mathbf{W}^{fb}$ for feedback connections are randomly initialized and fixed; only the output weights $\mathbf{W}^{out}$ are adapted during training.

In the reservoir we used around 100 to 200 neurons and set the connectivities for the weight matrices to 0.1 each, i.e. 90% of the connection weights are 0. This sparse connectivity allows the reservoir to develop more sophisticated and non-linear dynamics [9].

The input vector is 51-dimensional – one dimension for each feature mentioned in Section 3.2. We chose the following 4 output neurons:

1. A neuron that is 1 for positive, -1 for negative feedback and 0 otherwise.
2. A neuron to react on positive feedback with the value 1 and 0 otherwise.
3. A neuron to react in the same way to negative feedback.
4. A neuron to detect speech in general, i.e. to be 1 if there is speech and 0 if there isn't.

The 1st neuron represents the final output we want to produce. However, training the network to directly learn the output of this neuron turned out to be difficult – the output neuron preferred to react only with a positive or with a negative amplitude; it was not able to learn the differences of positive and negative speech. Introducing three other neurons, which all concentrate on one aspect we want to detect, allowed the 1st neuron to produce the desired output. We combined the outputs of the single neurons for a more stable output:

$$\textit{combined output} = \frac{\textit{neuron}_{speech} \cdot (\textit{neuron}_{posneg} + \textit{neuron}_{pos} - \textit{neuron}_{neg})}{2}$$

We used the C++ library *aureservoir* [8] as an implementation of our ESN. Due to the high sample rate and the large feature vectors the included batch training algorithm was not applicable. We, therefore, extended the library with an online training algorithm as proposed by Dai et al. [3]. The idea is to use the error between the reservoir output and the desired output for the weight update:

$$\mathbf{y}_{n+1} = \mathbf{x}_{n+1} \cdot \mathbf{W}_n^{out}$$

$$\mathbf{e}_{n+1} = \bar{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1}$$
$$\mathbf{W}^{out}_{n+1} = \mathbf{W}^{out}_{n} + \alpha \cdot (\mathbf{x}_{n+1})^T \cdot \mathbf{e}_{n+1} + \beta \cdot (\mathbf{x}_n)^T \cdot \mathbf{e}_n$$

where $\bar{\mathbf{y}}$ is the desired output, $\mathbf{y}$ the reservoir output and $\mathbf{e}$ the error. α and β are called the learning and the momentum gain respectively; they control how fast the weights adapt depending on the error of this time step and the time step before.

4 Results

To evaluate the performance of our ESN framework, we trained several networks on exclusive subsets of utterances. We first conducted speaker-dependent cross-validation, then tested generalization abilities. Afterwards we analyzed which features yield the best results and finally examined cultural dependencies.

4.1 Speaker-dependent Cross-validation

For each test subject and for both conditions ("na" and natural) we trained three networks (each on two thirds of the data) and then tested on the remaining third. Summing up the results we get the performance for this condition.

We tested with different thresholds for the network output amplitude. With a higher threshold a higher mean amplitude is needed as network output; if the threshold is not reached, the utterance is rejected. Therefore, we get higher accuracy at the cost of a higher rejection rate. The results, summed up over all test subjects, can be seen in Figure 3.

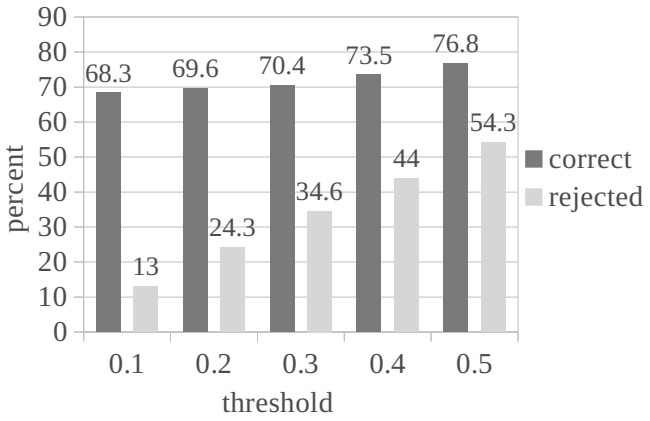

Figure 3: Percentage of correct classification for different thresholds.

When splitting the data into groups (defined by gender, culture and the type of feedback) the performances are spread as shown in Figure 4. The best performance is reached for the German females in the "na" condition with 81%. There is no significance in the differences between the performances of the natural and the "na" feedback, but we see a tendency that the classification on the "na" condition works better for female, while for male the natural condition yields better results. The Japanese in general perform slightly worse, but especially the Japanese female results have to be rated carefully, as there were considerably less data.

4.2 Looking at Generalization Capabilities

If the way in which the test subjects gave "na" feedback corresponds to the way they tend to give feedback in natural speech, it is expected that the "na" trained networks also work for the natural feedback data to some extend.

To test this, we used the three "na" trained networks per person to classify the natural feedback from this person and chose the best performing networks each. (Because of the randomly initiated reservoir of the network, it is expected that some might work better than others.) The overall classification accuracy is then 67.4%. The standard deviation is quite high, as for some subjects the classification did not work at all (yielding a classification rate around 50%). But

for two thirds of the test subjects the accuracy is higher than 60% and nearly half of the subjects perform even better than 70%. See Table 1 for the detailed results.

It can be observed for some people with very distinctive "na" feedback that we get better performance on the natural data classification when using "na" for training than when using the natural data for training. This leads to the assumption that an optimized network, trained only with the carefully chosen most distinctive utterances could improve the performances also for the evaluation in 4.1.

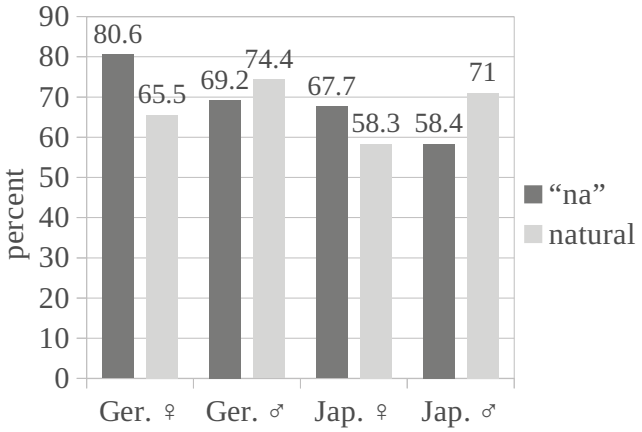

Figure 4: Percentage of correct classification in the different groups (with a fixed threshold of 0.3).

4.3 Feature Subset Analysis

Leaving out part of the features yields the results shown in Figure 5. Surprisingly, a combination of all features works only slightly better than MFCC features alone. Leaving out the MFCCs causes a significant drop in the overall performance, so they actually appear to be useful for the classification of prosody in this case.

The variance of the performance over several runs of the classifier is low, but the rejection rate varies strongly; it is most stable when all features are used.

4.4 Cross-cultural analysis

To account for cross-cultural differences we tested the networks trained on German data on other Germans and on Japanese data and vice versa. If there are differences in the way German and Japanese give feedback, we expect the German nets to perform better on other German data than on Japanese and the Japanese nets to perform better on Japanese audio data.

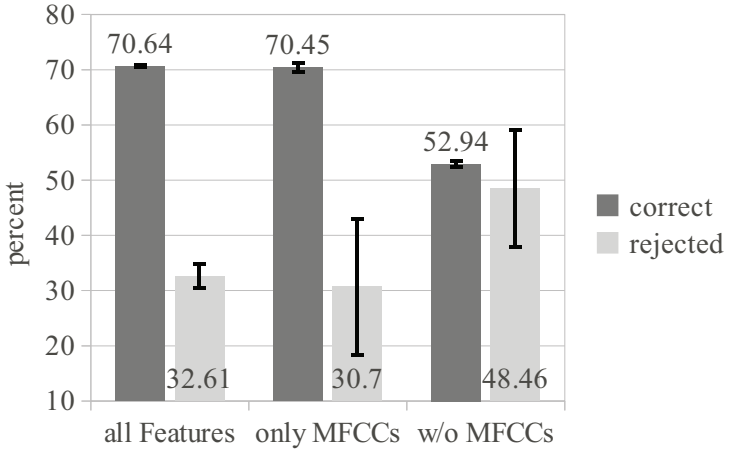

Figure 5: Performance of different features subsets with accuracy and rejection rate, averaged over 3 runs with all data.

The first assumption turns out to be true, as the German net performs better on German data for the "na" as well for as for the natural condition (see Figure 6). The difference is significant, but not that obvious, as also intra-cultural differences between the data are high. In general, there seem to exist very diverse strategies of giving feedback, even among people with the same cultural background.

For the Japanese networks a better performance on the same culture can only be seen in the natural data, but significance cannot be found here.

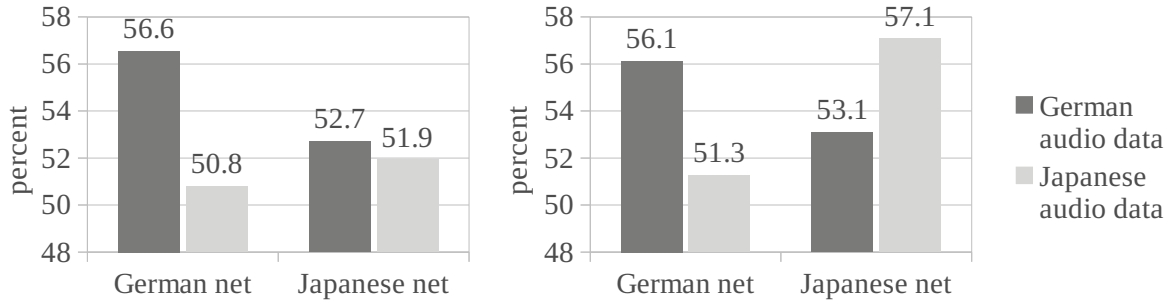

Figure 6: Results for ESNs applied to data of the same and of the other culture, on the left hand for "na", on the right hand for natural feedback.

5 Discussion

Our approach worked for "na" utterances as well as for utterances in natural language. From male persons the natural feedback could be recognized even better, while from females the "na" condition was easier to classify. Although not significant, these differences fit the experimenter's observations that most female persons easily used prosody in "na" feedback, while some males had difficulties to convey their intent without using words.

Table 1: Classification of natural data with "na" trained networks: Percentages of subjects that reach an accuracy higher than 50 / 60 / 70 / 80 %.

Mean performance	67%
Standard deviation	13.9
better than 50%	92%
better than 60%	67%
better than 70%	46%
better than 80%	21%

While some test subjects produced speech including high prosody, others relied more on lexical information using adult-directed speech. This could be prevented in further studies by developing the experimental design more carefully. For Flobi it has not been proved that he triggers infant-directed speech, so he might have been perceived by the participants in different ways. It might also be an idea for improvement to choose a more child-like voice for the robot's utterances.

The usage of prosody in feedback appears to be also highly situation dependent. Concerning the scenario, a task without any usage of lexical information might be better suited to enforce more emotional responses. Also a continuous task where the user's response directly influences the robot's next action could yield stronger prosody.

As the generalization properties seem to depend on the quality of the training data, training with data from actual infant-directed speech might be useful and could make it possible to also classify cases with more subtle prosody.

MFCCs are the standard approach for speech recognition, but are also often included additionally to the prosodic features for emotion recognition tasks. That MFCCs alone proved to work so well might be, because the MFCCs contain the whole speech signal, which can not be reconstructed through the prosodic features alone.

Looking back at our research questions we can conclude that differences between German and Japanese users seem to exist. A network trained on German data works significantly better on German audio data than on Japanese audio data, but a significant difference cannot be found the other way round. This could result from the fact that the prosody in Japanese was often very indistinct, at least in the explored group. The reason for that might be cultural influences, as a clear "No" is rarely used in Japanese everyday language.

Furthermore, it is possible to recognize positive and negative feedback from utterances in a speaker-dependent way. To some extent also speaker-independent networks would be possible to implement; but we observe very diverse strategies of giving feedback intra-culturally as well as inter-culturally. This indicated that a general solution is difficult to achieve and a speaker-dependent model for classification would be the better choice.

References

[1] BOTINIS, A., B. GRANSTRÖM and B. MÖBIUS: *Developments and paradigms in intonation research*. Speech Communication, 33(4):263–296, 2001.

[2] BREAZEAL, C. and L. ARYANANDA: *Recognition of affective communicative intent in robot-directed speech*. Autonomous robots, 12(1):83–104, 2002.

[3] DAI, J., G. VENAYAGAMOORTHY and R. HARLEY: *An introduction to the echo state network and its applications in power system*. In *ISAP'09. 15th International Conference on Intelligent System Applications to Power Systems, 2009*, pp. 1–7. IEEE, 2009.

[4] FERNALD, A.: *Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages*. Child development, 64(3):657–674, 1993.

[5] FERNALD, A., T. TAESCHNER, J. DUNN, M. PAPOUSEK, B. DE BOYSSON-BARDIES, I. FUKUI et al.: *A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants*. Journal of child language, 16(3):477–501, 1989.

[6] FRIEND, M.: *The transition from affective to linguistic meaning*. First Language, 21(63):219–243, 2001.

[7] FRIEND, M. and J. BRYANT: *A developmental lexical bias in the interpretation of discrepant messages*. Merrill-Palmer Quarterly: Journal of Developmental Psychology, 2000.

[8] HOLZMANN, G.: *Echo State Networks with Filter Neurons and a Delay&Sum Readout with Applications in Audio Signal Processing*. Master's thesis, Graz University of Technology, Austria, 2008. http://aureservoir.sourceforge.net.

[9] JAEGER, H.: *The "echo state" approach to analysing and training recurrent neural networks*. GMD Report 148, GMD - German National Research Institute for Computer Science, 2001.

[10] LUETKEBOHLE, I., F. HEGEL, S. SCHULZ, M. HACKEL, B. WREDE, S. WACHSMUTH and G. SAGERER: *The bielefeld anthropomorphic robot head "Flobi"*. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 3384–3391. IEEE, 2010.

[11] MORTON, J. and S. TREHUB: *Children's understanding of emotion in speech*. Child development, 72(3):834–843, 2003.

[12] SKOWRONSKI, M. and J. HARRIS: *Automatic speech recognition using a predictive echo state network classifier*. Neural networks, 20(3):414–423, 2007.

[13] THOMSON, D. and R. CHENGALVARAYAN: *Use of periodicity and jitter as speech recognition features*. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, vol. 1, pp. 21–24. IEEE, 1998.