EXPERIMENTS IN CONTEXT-INDEPENDENT RECOGNITION OF NON-LEXICAL ‘YES’ OR ‘NO’ RESPONSES

Shiva Sundaram, Robert Schleicher, Nathalie Diehl

Deutsche Telekom Laboratories, Ernst-Reuter-Platz-7, 10587 Berlin, Germany. {shiva.sundaram,robert.schleicher,nathalie.diehl}@telekom.de

ABSTRACT

We present our experiments in context-free recognition of non-lexical responses. Non-lexical verbal responses such as *mmm-hmm* or *uh-huh* are used by listeners to signal confirmation, uncertainty in understanding, agreement or disagreement in speech-based interaction between humans. Correct recognition of these utterances by speech interfaces can lead to a more natural interaction paradigm with computers. We present our study on both human and automatic recognition of positive (yes) and negative (no) non-lexical utterances. Approximately 3000 isolated utterances from 26 German native speakers were collected in a human-human spoken interaction setting. Our experiments indicate that human recognition accuracy is close to 99% and up to 90% accuracy can be obtained by standard automatic recognition techniques.

Index Terms— Speech interfaces, human-computer interaction, non-lexical responses, interjections, back-channel responses, automatic recognition, feature extraction.

1. INTRODUCTION

In addition to recognition of spoken words, non-lexical responses (NLRs) allow humans to interact in a natural way. Non-lexical utterances such as *mmm-hmm* or *uh-huh* are extensively used in dialogues (and monologues) to signal agreement, disagreement, confirmation (positive and negative) and uncertainty without relatively long, complex spoken lexical interaction. As the terminology suggests, in this study we define the scope of non-lexical responses as non-lexical (non word) sounds that lack elaborate semantic information. They are typically used to communicate limited information (such as yes, no, maybe, surprise, fear etc.) and may or may not be emotionally coloured. We are interested in its usage as positive confirmation (*yes*) or negation (*no*) outside dialogue situations which we take to be context-independent.

Non-lexical sounds cover a wide range of vocalized and non-vocalized sounds such as laughter, filled pauses, throat clearing, tongue clicks, lip smacks etc. In interpersonal communication they can appear as back-channels [1, 2], interjections [3] or even as filled pauses. Within the context of back-channels we focus on a subset of utterances which are not lexical and which are typically used for a positive confirmation or negation. We are particularly interested in the context-independent usage of NLRs. In applications such as an interactive voice response (IVR) system, and human robot interaction, isolated recognition of NLRs can be relevant for detecting a *yes* or *no* from a user. In more sophisticated settings, NLRs can also enable detecting confirmation or uncertainty in understanding (e.g. the options in a voice-driven menu system). As emotional variations such as disgust or pleasantness can also be imparted to these utterances through prosodic modifications, their analysis can also allow detection of the emotional state of a user.

In this paper, we present our initial study toward answering two main questions: (1) can humans recognize the specific *sense* of non-lexical responses without context (2) what acoustic features can enable automatic recognition of non-lexical responses. We describe our data collection procedure and report our baseline experiments in both human and automatic recognition of isolated non-lexical equivalents of *yes* or *no* responses. Performance of human recognition of these utterances will help establish its potential in identifying user state while interacting with machines. The underlying rationale being if humans themselves are unable to discern differences in the sense of NLRs, then it has little use for machine-based interaction. In a similar vein, the performance of automatic recognition established in our analysis will serve as a baseline of existing approaches and as a foundation for exploring future directions.

In the next section, we describe studies related to non-lexical utterances. Subsequently the main contributions of this work are highlighted.

1.1. Related Work

One of the broadest examinations of German non-lexical back-channels has to be considered the work by Ehlich [3], who does not use the term *back-channels*, but describes the phenomenon under the name of *interjections*. In his treatise different forms of German *hmm* responses within context were pragmatically analysed and compared to Chinese tone patterns. The author found differences in intonation to be characteristic for different meanings. Similarly, in [4], the
In comparison to the number of studies in spoken language processing on NLRs as back-channels [1, 2], there are a limited number of studies in isolated, context-independent setting. In [5], the authors studied the subjective perception and acceptability of a more generalized interpretation of non-lexical responses that is referred to as listener feedback. The analysis was strictly affective; the feedback responses belonged to ten different categories such as disgust, threat, boredom, relief, admiration etc. More recently, in [6] the authors study the relationship between physical acoustic properties and perceived characteristics of emotional non-lexical responses. The responses were particularly open-ended where the collected data comprised of affective vocalizations without a particular lexical character.

In relation to the taxonomy introduced in [7], in the current work, we only deal with simple utterances. In contrast to approaches presented in [8] that model dialogue acts or context-dependent information, in this paper we only focus on isolated NLRs that are used to answer ‘yes’ (a positive confirmation) or ‘no’ (a negation). Additionally, in contrast to [9, 6], we only consider responses that are neutral with regard to their emotional connotation.

2. DATA COLLECTION

We collected natural non-lexical responses of German native speakers (26 subjects; mean 25 years, 8 females) The speech was recorded in a sound-proof booth where the participants could verbally interact with a trained experimenter using a head mounted close-talk microphone (SHURE SM10a) and a headphone (Sennheiser 280HDPRO). The experimenter was seated outside the booth and he/she could also interact with the participant using the same microphone-headphone set up. They could also see each other through a transparent window. The overall set up was designed to keep the subject and experimenter acoustically isolated such that they could hear each other only through the audio feed from the recording equipment. Their speech were recorded in separate tracks simultaneously using a M-Audio MicroTrack-II stereo recorder at 48kHz. Each recording session with each subject lasted for 45 minutes and was composed of three parts.

The first introductory part was conducted to familiarize the subject with the recording session. The subject was encouraged to respond informally with an utterance indicating yes (positive) or indicating no (negation) without explicitly articulating ja or nein (yes or no in German). The second part of the session was also verbal and comprised of asking the subjects questions that were designed to elicit a ‘yes’ or a ‘no’ response. As no differences between replies to verbal and on-screen questions were noted during labelling, they were treated equally in our subsequent analysis. To avoid biased responses, care was taken to ask only neutral questions (such as favourite colour etc.) and not to ask any personal information or relating to topics such as politics or the environment. To establish ground truth for a response being a yes or a no, the experimenter also noted down the interpretation of the response on paper based on facial expression, head nods and the question posed during all parts of the session. For instance, if the weather outside was deemed to be warm, the question posed to the participant was “Is today a cold winter day?” and the response was noted as a ‘no’. Questions that could lead to ambiguous or uncertain responses were also mixed in, but these were not considered in the subsequent analysis. The third part of the session was completed using a computer monitor that was set up inside the booth but was controlled by the experimenter. The computer monitor was used to present (in text) one question at a time. The questions were similar in spirit to the second part but were not asked verbally by the experimenter. The responses from the second and third part of the session was the data mainly used in this paper.

2.1. Data preparation and analysis

First, the audio data from 26 speakers were down sampled to 16 kHz. The recordings from the left channel (the participants’ speech) were then manually segmented using WaveSurfer [10]. With the experimenter’s notes as the reference, the segmented responses were labelled as a yes, or as a no. In all, the 3000 (Y/N=1518/1482) isolated utterances were available for subsequent analysis. A subset of about 50% of these were used for the subjective classification experiments and 3000 were used for the automatic classification experiments. This is described next.

3. EXPERIMENTS

3.1. Listening Experiments

To estimate the ability of humans to detect a ‘yes’ or a ‘no’ in context-independent NLRs, we performed an independent subjective listening experiment. 59 subjects (none from the data collection) participated in the listening experiments (36 male, mean age 24.2 years) that were conducted in an acoustically treated room. 1446 (roughly 50%) of the utterances marked as clear yes or no were split into sets of approximately 300 clips. These clips were played over headphones through a computer using a web-based interface. The subjects were instructed to rate each clip according to their perceived impression of yes and no. There ratings were performed on a five-point scale that ranged from ‘clear yes’ to ‘clear no’ with intermediate steps ‘rather yes/no’ and ‘don’t know’ in case of indecision. Even though these values were used in a binary fashion later on, in our experience, offering intermediate steps clearly decreases the number of ‘don’t know’ answers. The listening experiment was conducted in multiple sessions such that each set was rated by at least five subjects. Each session lasted for about 45 - 80 mins.
Table 1. Signal Features used for automatic classification experiments.

<table>
<thead>
<tr>
<th>Type</th>
<th>Features (Abbreviation, Dimensionality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame-level</td>
<td>pitch (P, 1), $\delta P^2$ (DP, 1), $\delta^2 P$ (D$^2$P, 1), short term energy (E, 1), $\delta E^2$ (DE, 1), Mel-freq. Cepstral Coeff. (MFCC, 13), $\delta (MFCC)$ (DMFCC, 13), $\delta^2 (MFCC)$</td>
</tr>
<tr>
<td>Global</td>
<td>mean(P)(Pmin, 1), std(P)(Pstd, 1), min(P)(Pmin, 1), max(P)(Pmax, 1), mean(E)(Emn, 1), std(E)(Estd, 1), min(E)(Emin, 1), max(E)(Emax, 1)</td>
</tr>
</tbody>
</table>

Table 2. Subjective classification confusion matrix. 1446 utterances in total. Chance-level approximately 50%.

### 3.2. Automatic Classification

Automatic classification experiments were performed using standard approaches to distinguish between isolated instances of ‘yes’ and ‘no’ back-channel responses. Two types of classifiers were trained and tested: a hidden Markov model based (HMM) and a Gaussian mixture model (GMM) based classifier. Features extracted frame-wise were used for the HMM-based classifier and averaged utterance-level features were used for the GMM classifier. Details of the signal features is shown in table 1.

**HMM-based Classifier:** A left-right HMM each for yes and no utterances was trained using the Baum-Welch algorithm. The label of the best model that explained the observations from a given test utterance in the Viterbi decoding step was used as the predicted label. During training, the individual utterances were used without any context-dependent information. Inspired by its application in speech recognition, the combination of frame-level features and HMM model the overall trajectory of the utterance using the cepstral and prosodic features.

**GMM-based Classifier:** One Gaussian mixture model (GMM) model each for yes and no back-channel utterances was trained using the Expectation-Maximization (EM) algorithm. The GMMs were trained on a reduced dimensional space resulting from principal component analysis (PCA) of the global features. Experiments with different combinations of the signal features were performed.

### 4. RESULTS

#### 4.1. Subjective Classification Results

The individual ratings from section 3.1 were averaged for each utterance across all listeners and this mean value then discretized to obtain a binary ‘yes’ or ‘no’ rating for a given clip. Table 2 shows the classification according to naive listeners compared with the labelling of our expert annotator. Altogether there were only 4 (=0.3%) and 6 (=0.4%) ‘no’ and ‘yes’ mis-classifications out of the 1446 samples that had to be judged by naive participants. As the 10 misclassified stimuli were from samples from nine different speakers, there was no specific pattern apparent like for instance, utterances of a specific person’s utterances being ambiguous. High subjective classification performance of isolated utterances indicates that the acoustic information of ‘yes’ and ‘no’ were sufficiently distinct and well formed. Additionally, considering the high agreement between naive listeners who did not have access to other information except the audio signal, and expert labelling that also took into account cues like nodding (see section 2.1), only 4 respective 6 mis-classifications out of 1446 samples validated our ground truth labels during data annotation.

#### 4.2. Results of Automatic Classification

Various performance measures for different classifiers and feature set combinations were estimated by a speaker-wise leave-one-out procedure. In each fold, the utterances from one speaker were retained as the test set and the training set was comprised of the utterances from the remaining 26 speakers. The performance metrics were the average of the results from the 26 leave-one-out folds.

**HMM-based Classifier:** Eight meaningful combinations or early-fusion of cepstral and prosodic information at the frame-level were tested with the HMM-based classifier. The five best performing classifiers are summarized in table 3. For the feature sets that included MFCCs, an 8-state 4 mixture per state HMM and for the models that only used the pitch and energy measures, a 5-state 4 mixture per state model were empirically determined to perform the best.

**GMM-based Classifier:** The performance measures for the seven best performing feature set combinations with GMM classifiers are given in table 3. In each case, 3 to 4 mixture components (with shared covariance) were empirically determined to give the best results.
<table>
<thead>
<tr>
<th>Feature Set (Dimensionality)</th>
<th>% P</th>
<th>% R</th>
<th>% F</th>
<th>% A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human classification</td>
<td>99.55</td>
<td>99.18</td>
<td>99.31</td>
<td>99.31</td>
</tr>
<tr>
<td>Frame-level Features + HMM-based classification performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC + DMFCC + D^2MFCC (39)</td>
<td>88.95</td>
<td>88.86</td>
<td>88.88</td>
<td>90.66</td>
</tr>
<tr>
<td>MFCC + DMFCC + D^2MFCC + P + DP + D^2P (42)</td>
<td>88.06</td>
<td>88.09</td>
<td>88.06</td>
<td>90.30</td>
</tr>
<tr>
<td>P + DP + D^2P + E + DE + D^2E (6)</td>
<td>84.77</td>
<td>85.02</td>
<td>84.89</td>
<td>87.00</td>
</tr>
<tr>
<td>P + E (2)</td>
<td>81.68</td>
<td>82.10</td>
<td>81.82</td>
<td>83.50</td>
</tr>
<tr>
<td>P + DP + D^2P (3)</td>
<td>82.13</td>
<td>82.00</td>
<td>82.04</td>
<td>82.30</td>
</tr>
<tr>
<td>Global-features + GMM-based classification after PCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pmn + Pmn + Pmax + Pstd + Emm + Emin + Emax + Estd (8)</td>
<td>71.12</td>
<td>71.06</td>
<td>71.09</td>
<td>71.10</td>
</tr>
<tr>
<td>Pmn + Pstd (2)</td>
<td>68.27</td>
<td>68.25</td>
<td>68.26</td>
<td>68.27</td>
</tr>
<tr>
<td>Pmn + Pmax + Pstd (4)</td>
<td>66.75</td>
<td>66.65</td>
<td>66.70</td>
<td>66.70</td>
</tr>
<tr>
<td>Pmn + Pmax (2)</td>
<td>66.42</td>
<td>66.41</td>
<td>66.42</td>
<td>66.43</td>
</tr>
<tr>
<td>Emm + Emin + Emax + Estd (4)</td>
<td>66.10</td>
<td>66.04</td>
<td>66.07</td>
<td>66.00</td>
</tr>
<tr>
<td>Emm + Estd (2)</td>
<td>63.53</td>
<td>63.40</td>
<td>63.46</td>
<td>63.34</td>
</tr>
<tr>
<td>Emin + Emax (2)</td>
<td>57.15</td>
<td>57.14</td>
<td>57.14</td>
<td>57.11</td>
</tr>
</tbody>
</table>

Table 3. Average performance measures (P-Precision, R-Recall, F-F measure, A-Accuracy) of different classifiers. Dashed lines indicate that the differences in prediction performance were not statistically significant in pairwise comparisons in Analysis of Variance (ANOVA) (p ≤ 0.05, Sidak-corrected for multiple comparisons).

5. CONCLUSION

In this work, we presented our experiments in context-independent classification of non-lexical ‘yes’ and ‘no’ responses. Both subjective and automatic classification experiments were performed and their respective results were presented. Related procedure to collect isolated back-channel utterances in natural human-human communication was also described.

Our experiments and related analysis strongly support our main hypotheses. First, from the subjective experiments, it was found that non-lexical responses contain enough acoustical information for positive confirmation and negation to be correctly identified by humans without additional context information (section 4.1). Therefore, they have potential for application in speech-based interfaces to allow for more natural interaction.

Second, experiments using frame-level and utterance-level features suggest strong evidence of both time-varying spectral information (similar to lexical, spoken words) and globally distinct prosodic information that can be exploited for automatic classification. In this respect, although the frame-level features showed superior classification performance to utterance-level features, both can potentially be combined to improve the overall accuracy and be at par with human recognition performance. Our results (summarized in Table 3) also show that the performance of frame-level pitch and energy features is comparable with frame-level cepstral features (rows 1, 2 and 3). This indicates that prosodic features contain sufficient information for distinguishing the sense of the responses. Thus supporting Ehlich’s [3] tonal analysis and Werner’s [4] findings of functional dichotomy in pitch profiles. While our initial experiments in direct combination of cepstral and prosodic features (such as presented here) did not show significant improvement, we believe a more sophisticated approach of combining frame-level and utterance-level measures could indeed show significant gains. This is a part of our current and planned future work.

In addition to this, we will further explore the gamut of information available in non-lexical responses. This includes analysis of utterances that serve to signal uncertainty in understanding or apprehensiveness in agreement/disagreement and also utterances with emotional connotation such as disgust or pleasantness. Not only will this help with our original motivation of improving natural interaction with computers, it will also help communicating a variety of information from human to machines (and vice versa) clearly and naturally without long lexical/spoken interaction. A meaningful start to decoding and discerning this can be through analysis of syllabification [1] or even phonetic segmentation of non-lexical utterances.

6. REFERENCES