Automatic Detection and Evaluation of Edentulous Speakers with Insufficient Dentures

Tobias Bocklet[1,2], Florian Hönig[1], Tino Haderlein[1,3], Florian Stelzle[2], Christian Knipfer[2], and Elmar Nöth[1]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstraße 3, 91058 Erlangen, Germany
tobias.bocklet@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

[2] Universität Erlangen-Nürnberg, Mund-, Kiefer- und Gesichtschirurgische Klinik, Glückstraße 11, 91054 Erlangen, Germany

[3] Universität Erlangen-Nürnberg, Abteilung für Phoniatrie und Pädaudiologie, Bohlenplatz 21, 91054 Erlangen, Germany

Abstract. Dental rehabilitation by complete dentures is a state-of-the-art approach to improve functional aspects of the oral cavity of edentulous patients. It is important to assure that these dentures have a sufficient fit. We introduce a dataset of 13 edentulous patients that have been recorded with and without complete dentures in situ. These patients have been rated an insufficient fit of their dentures, so that additional (sufficient) dentures and additional speech recordings have been prepared. In this paper we show that sufficient dentures increase the performance of an ASR system by ca. 27 %. Based on these results, we present and discuss three different systems that automatically determine whether the dentures of an edentulous person have a sufficient fit or not. The system with the best performance models the recordings by GMMs and uses the mean vectors of these GMMs as features in an SVM. With this system we were able to achieve a recognition rate of 80 %.

Key words: speech recognition, user modeling, assistive technology, applied system

1 Introduction

A complete loss of teeth can cause persistent speech disorders by altering dental articulation areas. This severely reduces the quality of speech [1] and even the recognition rate of an automatic speech recognition system [2,3]. Removable complete dentures partly solve these problems [4].

Dental rehabilitation of edentulous patients by complete dentures improves not only esthetic and functional aspects, e.g., mastication of food, but also the speech quality. However, complete dentures restrict the flexibility of the tongue, narrow the oral cavity and alter the articulation areas of the palate and teeth. This is even amplified when the fit of the dentures is not sufficient.

In this paper we first introduce a dataset of 13 edentulous speakers that have been recorded with and without complete dentures in situ. After a later examination these

Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): TSD 2010, LNAI 6231, pp. 237–244, 2010.
© Springer-Verlag Berlin Heidelberg 2010

dentures have been rated not to have a perfect fit, i.e., an insufficient fit. This aspect required a new preparation of a complete denture for each of these 13 patients. So finally three recordings where available for each of these 13 patients: one without complete dentures, one with insufficient dentures and one with sufficient dentures.

In this paper we first examined in which way these three types of recordings affect the performance of an automatic speech recognition system. The main goal of this paper was to create a system that is able to detect on a spoken text whether the complete dentures have a sufficient fit or not. We describe three different systems to achieve this goal: One system takes the word accuracy results of a speech recognizer and uses this single value as feature. The second system calculates the distance between the spoken text of an edentulous patient and a reference speaker by Dynamic Time Warping (DTW). These distances are then used as features for a classifier. The other system models each speaker by Gaussian Mixture Models (GMMs) and uses the concatenated mean vectors of these GMMs as input vector for a classifier.

The outline of this paper is as follows: Section 2 describes the dataset we used, Section 3 describes our recognition system and the two systems we used for detecting insufficient dentures. The results are presented and discussed in Section 4. We finish with a summary in Section 5.

2 Dataset

The original dataset was first introduced in [2]. In contains 28 edentulous, i.e. toothless, patients. Their average age was 64 ± 10 years. Only patients who wore removable complete dentures where chosen to participate in this study. Only patients wearing their dentures for at least one month where chosen in order to ensure a patient habituation to their new dentures. All patients were native German speakers who were asked to speak standard German while being recorded. None of the patients had speech disorders caused by medical problems others than dental or any report of hearing impairment.

After a later examination by a senior dentist 13 patients had been rated to have complete dentures with an insufficient fit. The complete denture was rated as insufficient if one of the following seven parameters was rated as insufficient/ not correct:

- Absence of pain concerning the chewing muscles, the soft and hard tissue in functional and non- functional situations
- Absence of variances of the soft tissue like redness or ulcer
- Ability to chew and swallow without restrictions
- Balanced occlusion relationship under function
- An interocclusal distance of 2 mm in a physiologic rest position
- Excellent fit proven by a soft pattern
- Patient satisfaction

For these 13 patients, additional dentures have been produced which have been rated a sufficient fit afterwards. Finally, the patients read the text “Der Nordwind und die Sonne” (NWS) three different times: One time without their complete dentures in situ, one time while wearing their insufficient dentures and one time wearing their sufficient dentures. The NWS text is a phonetically balanced text with 108 words (71 disjunctive) which is

used in German speaking countries in speech evaluation and therapy. The speech data was sampled with a frequency of 16 kHz and an amplitude resolution of 16 bit.

3 Methods

In this section we first describe our feature front-end processing. These features are used in our ASR system (Section 3.2) and in two of the three denture classification systems (see Section 3.3).

3.1 Feature Extraction

As features we use the well-known Mel-frequency cepstrum coefficients (MFCCs). These features perform a short-time analysis of the speech signal. Therefore a Hamming window with a length of 16 ms and a frame rate of 10 ms is applied to the signal. The filterbank for the Mel spectrum with 25 triangle filters is calculated afterwards. The cepstral coefficients are computed by an inverse discrete cosine transform of the logarithmic Mel spectrum. For each frame a 24-dimensional feature vector is created. It contains the short-time energy, 11 Mel-frequency cepstral coefficients and their first-order derivatives approximated by the slope of a regression over 5 consecutive frames.

3.2 Automatic Speech Recognition System

The speech recognition system used for some of the experiments was developed at the Chair of Pattern Recognition in Erlangen [5]. The system is based on semi-continuous Hidden Markov Models (HMM). It can model phones in a context as large as statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones. The HMMs for each polyphone have three to four states with a codebook containing 500 Gaussian mixtures with full covariance matrices. The feature front-end of our ASR system is described in Section 3.1.

Our ASR system was trained on German dialogues from the VERBMOBIL project [6]. The data was recorded with a close-talking microphone and a sampling frequency of 16 kHz. It was quantized with 16 bit. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. 11,714 utterances (257,810 words) of the VERBMOBIL-German data (12,030 utterances, 263,633 words, 27.7 hours of speech) were used for training and 48 (1,042 words) for the validation set, i.e. the corpus partitions were the same as in [5].

The vocabulary of known words of the ASR system was changed to the 71 words of the NWS text. The word accuracy (WA) and the word recognition rate (WR) are used as basic measures. They are computed from the comparison between the recognized word sequence and the reference text consisting of the $n_{\text{all}} = 108$ words of the read text. With the number of words that were wrongly substituted (n_{sub}), deleted (n_{del}) and inserted (n_{ins}) by the recognizer, the word accuracy in percent is given as

$$\text{WA} = [1 - (n_{\text{sub}} + n_{\text{del}} + n_{\text{ins}})/n_{\text{all}}] \cdot 100 \tag{1}$$

Only a unigram language model was used in our recognition experiments in order to put more emphasis onto the acoustic models.

3.3 Classification Systems

In this section we describe the three systems we proposed to detect insufficient dentures automatically. Our baseline system is based on the word accuracy rates of the speech recognizer. The second system uses dynamic time warping (DTW) and calculates the distance between a speaker with dentures and a reference speaker. The third system models each speaker with Gaussian Mixture Models (GMMs) and uses the GMMs as meta-features within a classification system.

System with word accuracy features Our baseline system uses a one dimensional feature vector. This feature represents the WA of the speech recognition system as single feature. The system is motivated by the fact, that we achieved different mean WA values for the three subsets of our dataset (see Section 4.1). This one-dimensional feature vector is then classified by a Support Vector Machine (SVM) [7].

System with Dynamic Time Warping In a second approach, we use Dynamic Time Warping (DTW) [8] to extract a feature vector coding distances between the recording of the speaker with dentures to be assessed (*test speaker*) and a (non-pathologic) reference realization of the same text in the MFCC feature space. This distance feature vector is then classified by an SVM.

More precisely, we compute the series of acoustic feature vectors for the two realizations of the NWS – text (mean-normalized MFCCs, augmented by the first derivatives), and compute a mapping between them such that the accumulated (squared) distance between corresponding feature vectors plus the costs incurred by insertions and deletions is minimal (we use a suitable weighting of energy (first cepstral coefficient), the derivatives and insertions/deletion penalties). This mapping relates corresponding phonemes of the test and reference speaker very reliably since the recordings contain the same words; where the test speaker inserts or deletes words, according large insertion/deletion penalties occur. Substituted words or phonemes are usually reflected by both larger acoustic distances and higher insertion/deletions penalties. We then construct a feature vector for classification that contains for each frame j of the reference sound the average distance to those acoustic feature vectors of the test speaker that frame j is assigned to, and the penalties for insertions and deletions. Thus, we obtain a feature vector of fixed size (which is necessary for classification) that details the test speaker's deviations from the reference speaker over time. It is left to the classifier to concentrate on those parts of the recording (i.e. elements of the distance feature vector) that are important for classification performance. This strategy is much more promising than e.g. just using the average distance between assigned feature vectors.

System with Gaussian Mixture Models The third system is based on Gaussian Mixture Models (GMMs). We used this system for the task of intelligibility assessment on children with Cleft Lip and Palate [9] and on speakers with partial laryngectomy [10]. We base our evaluation only on a shift of the mean vectors from a Universal Background Model (UBM) to the mean vectors of a speaker GMM. GMMs model the acoustic features, and with this the acoustic space, of a specific recording. A GMM (λ) contains

M unimodal Gaussian densities. Each density represents a different acoustic area of the feature space:

$$p(\vec{c}|\vec{\lambda}) = \sum_{i=1}^{M} \omega_i \, p_i(\vec{c}|\vec{\mu}_i, \vec{\Sigma}_i) = \tag{2}$$

$$= \sum_{i=1}^{M} \omega_i \cdot \frac{1}{(2\pi)^{D/2}|\vec{\Sigma}_i|^{(1/2)}} e^{-(1/2)(\vec{c}-\vec{\mu}_i)^T \vec{\Sigma}_i{}^{-1}(\vec{c}-\vec{\mu}_i)}, \tag{3}$$

The idea is now to train GMMs, extract the mean vectors of the GMMs for each recording of one speaker, i.e., with sufficient and with insufficient dentures, concatenate them and classify these vectors with an SVM.

After feature extraction (Section 3.1) a UBM is created on a dataset of healthy speakers who also read the NWS text. This is achieved by using 5 iterations of the EM algorithm [11]. Beginning with this UBM, a speaker-dependent GMM is built by MAP adaptation [12]. The MAP adaptation takes the UBM as an initial model and adapts the statistics to the acoustic features of a specific speaker in a single iteration step. These new densities are combined with the UBM statistics afterwards. Finally, a GMM λ is created for each recording. The components of each GMM are concatenated to a GMM-based supervector. These supervectors can be regarded as a mapping from the acoustics of a recording, i.e., MFCCs, to a higher-dimensional feature vector which represents the acoustic characteristics of this recording. Since we are dealing with two types of recordings, i.e., recordings of persons with sufficient and insufficient dentures, we expect the acoustic characteristics of these two types being different and expect them to be modeled by these GMM supervectors.

4 Experiments and Results

The results of this paper can be split into two different parts: First we describe the WA results on the three different types of recordings of our 13 patients. These different types are recordings were the persons did not wear their dentures, recordings were they wear insufficient dentures, and recordings with sufficient dentures in situ.

In the second part of this section we present and discuss the results on recognizing whether a recording was performed with sufficient or insufficient dentures.

4.1 Speech Recognition

The automatically computed WA differed for the three different subsets (see Table 1). The WA on recordings without dentures (60.06 %) was lower than recordings with insufficient dentures (64.35 %). In the case of sufficient dentures a WA of 70.91 % was measured. This is an improvement by 18 % compared to the recordings without dentures and 10 % compared to the recordings with insufficient dentures. The standard deviation on recordings without any dentures was $\pm$10.35. The value decreased to $\pm$9.64 and $\pm$6.04 when wearing insufficient dentures or sufficient dentures, respectively. This is an improvement of 37 % when comparing the values of insufficient and sufficient dentures.

Table 1. Word accuracy (WA) result, according standard deviation and minimum/maximum WA value for the three different subsets: Without wearing dentures, with insufficient dentures, with sufficient dentures

dataset	mean WA	std. dev	min WA	max WA
without	60.06	± 10.35	39.48	77.78
insufficient	64.35	± 9.64	45.37	80.56
sufficient	70.91	± 6.04	57.51	80.56

This improvement is also visible when focusing on the minimum word accuracy values. Insufficient dentures improved the minimal WA from 39.48 % to 45.37 %. This is an improvement by 15 %. Wearing sufficient dentures again improved the value by 27 % to a WA of 57.51 %.

4.2 Denture Classification

The two-class problem of identifying whether the complete dentures have a sufficient fit or not was handled by three different systems. Since it was not the goal of this paper to select the classifier with the best performance, we selected an SVM to be used for each experiment. The dataset used in this paper had a very limited number of samples. It contains 13 speakers, with two recordings for each speaker; one time with insufficient dentures and one time with sufficient dentures. To deal with that problem, we performed our experiments in a cross-validation with leave-one-speaker-out manner.

System number one uses a one-dimensional feature vector, i.e., the word accuracy of the recognizer. With this baseline system, a recognition result of $61.5\dot{}$ % was achieved for the two-class problem; 7 recordings of sufficient dentures have been classified correctly, and 9 recordings with insufficient dentures have been classified correctly.

The second system uses the DTW distances as features for an SVM classification. These distances have been calculated with respect to the recordings of a reference speaker without dentures. The length of the feature vector was 2, 314. 8 of 13 insufficient recordings have been identified correctly and 11 recordings have been correctly identified as insufficient. This system achieved a recognition result of 73.1 %. This is a relative improvement of 19 % compared to the baseline system.

The third system uses the mean vectors of a 128-dimensional GMM as features. The dimension of this GMM supervector is $24 * 128 = 3,072$. Again, an SVM was used for classification. The number of correctly classified recordings with sufficient dentures was 10; 11 recordings of insufficient dentures have been classified correctly. This sums up to a recall of 80.1 %. Compared to the baseline system the GMM system achieved an significant improvement ($p < 0.1$) of 30 %. Compared to the DTW-based system the system achieved a relative improvement of 9.5 %.

5 Summary

In this paper we performed ASR experiments on a dataset of 13 edentulous patients. Complete dentures have been produced for these speakers. since these dentures have

Table 2. Recognition results of the three different systems on the problem of detecting insufficient dentures

system	feature dim	corr. sufficient	corr. insufficient	recognition result
WA	1	7	9	61.5 %
DTW	2,314	8	11	73.1 %
GMM	3,072	10	11	80.1 %

been rated an insufficient fit, new dentures have been created for these speakers. So for each speaker three recordings have been available. We performed ASR experiments on these data that showed an improvement of the mean WA of 10 % between recordings without any dentures and recordings with insufficient dentures. Sufficient dentures improved these results by another 18 %. The second task of this paper was the automatic identification of incomplete dentures. Therefore we compared the recognition results of three different system: one system used only the WA result as feature, a second system used the DTW distances w.r.t. to a reference speaker and the third system used the mean vectors of 128 dimensional GMMs as features. We achieved a recognition rate of 80 % for this two-class problem.

Acknowledgement

This work was supported by the Wilhelm Sander-Foundation, Germany (AZ: 2007.100.1). The authors are responsible for the content of this article.

References

1. Ichikawa, J., Komoda, J., Horiuchi, M., Matsumoto, N.: Influence of Alterations in the Oral Environment on Speech Production. Journal Of Oral Rehabilitation **22** (1995) 295–299
2. Haderlein, T., Bocklet, T., Maier, A., Nöth, E., Knipfer, C., Stelzle, F.: Objective vs. Subjective Evaluation of Speakers with and without Complete Dentures. In Matousek, V., Mautner, P., (eds.): Proc. Text, Speech and Dialogue; 12th International Conference. Volume 1 of Lecture Notes in Artificial Intelligence., Berlin (2009) 170–177
3. Stelzle, F., Uginovic, B., Knipfer, C., Bocklet, T., Nöth, E., Schuster, M., Eitner, S., Seiss, M., Nkenke, E.: Automatic, Computer-Based Speech Assessment on Edentulous Patients with and without Complete Dentures – Preliminary Results. Journal Of Oral Rehabilitation **37** (2010) 209–216
4. Tanaka, H.: Speech Patterns of Edentulous Patients and Morphology of the Palate in Relation to Phonetics. The Journal of Prosthetic Dentistry **29** (1973) 16–28
5. Stemmer, G.: Modeling Variability in Speech Recognition. Volume 19 of Studien zur Mustererkennung. Logos Verlag, Berlin (2005)
6. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)
7. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery **2**(2) (1998) 121–167
8. Sakoe, H., Chiba, S.: Dynamic Programming Algorithm Optimization for Spoken Word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **26**(1) (1978) 43–49

9. Bocklet, T., Maier, A., Riedhammer, K., Nöth, E.: Towards a Language-independent Intelligibility Assessment of Children with Cleft Lip and Palate. In: Workshop on Child, Computer, and Interaction 2009, New York (2009)
10. Bocklet, T., Haderlein, T., Hönig, F., Rosanowski, F., Nöth, E.: Evaluation And Assessment Of Speech Intelligibility On Pathologic Voices Based Upon Acoustic Speaker Models. In: Proceedings of the 3rd Advanced Voice Function Assessment International Workshop, Madrid (2009) 89–92
11. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological) **39**(1) (1977) 1–38
12. Gauvain, J., Lee, C.: Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing **2** (1994) 291–298