In “Speech Recognition and Coding — New Advances and Trends”, Rubio–Ayuso and Lopez–Soler (eds.), Springer, 1995, pp. 329–332.

1

Voice Source State as a Source of Information in Speech Recognition: Detection of Laryngealizations

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

ABSTRACT Laryngealizations are irregular voiced portions of speech, which can have morpho-syntactic functions and can disturb the automatic computation of F0. Two methods for the automatic detection of laryngealizations are described in this paper: With a Gaussian classifier using spectral and cepstral features a recognition rate of 80% (false alarm rate of 8%) could be achieved. As an alternative a “non-standard” method has been developed: an artificial neural network (ANN) was used for non-linear inverse filtering of speech signals. The inversely filtered signal was directly used as input for another ANN, which was trained to detect laryngealizations. In preliminary experiments we obtained a recognition rate of 65% (12% false alarms).

1.1 Introduction

The term “laryngealization” is used by us as a cover term for irregular, voiced stretches of speech, often accompanied by an extremely low pitch, that may occur inside one phone but can extend across phone boundaries as well. Other terms found in the literature are: vocal fry, creaky voice, pulse register, creak, etc. Usually laryngealizations do not disturb pitch perception but are perceived as suprasegmental irritations modulated onto the pitch curve. These special voice source phenomena occur frequently at distinct positions, e.g. at morpheme boundaries as in *“she eats”* vs. *“sheets”*. They have not been investigated very often, but have mostly been considered to be an irritating phenomenon that has to be discarded.

For pitch determination algorithms, laryngealizations cause severe errors [4, p. 49], and should therefore be localized and treated in a special way. In the past, feature extraction for speech recognition concentrated on articulatory information (mel-cepstral coefficients etc.), whereas voice source information (laryngeal information) has been almost totally neglected. Recently, the functionality of laryngealizations as, e.g. boundary markers has been noticed [5] [6]. In our material, e.g. 58% of all the laryngealizations were found at the beginning and 18% at the end of a word, i.e. $3/4$ of all laryngealizations were located at word boundaries. Knowledge about the location of laryngealizations could thus be used to improve pitch determination as well as speech recognition and parsing.

[1] Univ. Erlangen-Nürnberg, Lehrstuhl f. Mustererkennung (Inf. 5), 91058 Erlangen, F.R. of Germany

[2] L.M.-Univ. München, Institut für Deutsche Philologie, 80799 München, F.R. of Germany

[2] 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.

1.2 Database and Feature Extraction

For our experiments a database of 1329 sentences from 4 speakers (3 female, 1 male) was used (30 minutes of speech in total). For more details see [2]. For the whole database the laryngealized frames (12.8 msec each) were labeled by two phoneticians [1]. 4.8% of the speech in total (7.4% of the voiced speech) is laryngealized (1191 laryngealized portions of speech), which is comparable to other databases [6].

In the time domain laryngealizations are characterized by peculiarities that often show up clearly in the signal (cf. figure 1.1) as e.g. irregular periodicity, strong variations of the amplitude, special form of the damped wave, or very long pitch periods [5] [1]. In the spectrum it has been claimed that laryngealizations can be described as having a characteristic spectral tilt [7]. Therefore we investigated the use of frequency and time domain features for the detection of laryngealizations.

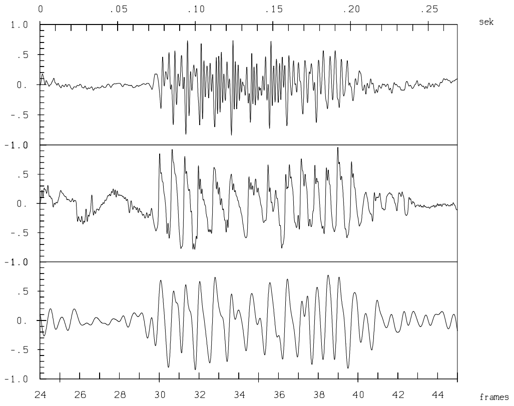

FIGURE 1.1: Speech signal (top) with laryngealization (frames 32 to 38), the corresponding laryngogramm (middle) and the glottis signal automatically computed by an artificial neural network (bottom).

1.2.1 Approach I (Frequency Features):

In preliminary examinations we noticed that the spectra and the cepstra of laryngealized phones and of their 'normally' phonated counterparts can differ in several ways. E.g. in figure 1.2 (bottom) the cepstrally smoothed spectrum of the non-laryngealized phone (right) shows a more regular harmonic structure than the spectrum of the laryngealized phone (left). We derived a set of potential features from the spectrum, the cepstrally smoothed spectrum and the cepstrum (for more detail cf. [8]). In a first step, the ability of the different features to discriminate between laryngealized and non-laryngealized phones was tested using Gaussian classifiers. Best results were achieved with 5 features: (1) the sum of the vertical distances of neighboring extrema and the average vertical distance of these extrema in the cepstrally smoothed spectrum below 1700 Hz, (2) the location and height of the absolute maximum in the cepstrum, (3) the quotient of the largest and the second largest maximum in the cepstrum of the center-clipped signal (to eliminate the influence of the vocal tract). In a second step, these 5 features were combined with normal mel-cepstral coefficients to train a phone component recognition system. Normally this system distinguishes between 40 different phones using 11 mel-cepstral coefficients per frame, a Gaussian classifier, automatic clustering into 5 clusters per phone and a full covariance matrix. Here we describe experiments with learn=test (speaker-dependent and multi-speaker)[3].

[3] Experiments with learn≠test (speaker-independent, leave-one-out). showed the same tendencies.

For each phone for which more than 100 laryngealized frames occur in our database (11 phones in total, in order of frequency: [a], [3], [aU], [e], [o], [I], [6], [U], [@], [n], [O]) a new additional phone label was introduced. By using only the 11 cepstral coefficients a recognition rate of 25% for the now 51 phones was achieved. By adding the 5 laryngeal features to the feature vector the recognition rate went up to 33%. Note, that the training database is very small, but we wanted to find out how well we can detect laryngealizations; thus the improvement is more important than the actual recognition rate. A mapping of the 40 non-laryngealized phones into one class, and the other 11 laryngealized phones into another class, yielded a recognition rate for laryngealizations of 80% with a false alarm rate of 8%. For leave-one-out (using 3 speakers for training, the other for testing) we obtained a recognition rate of 67% with a false alarm rate of 7%. Other experiments showed that this decrease in performance is due to a strong speaker-dependence of laryngealizations. Just using the 11 cepstral coefficients, 25% laryngealizations are found with a false alarm rate of only 1% showing that the cepstral coefficients already contain some information about laryngealizations.

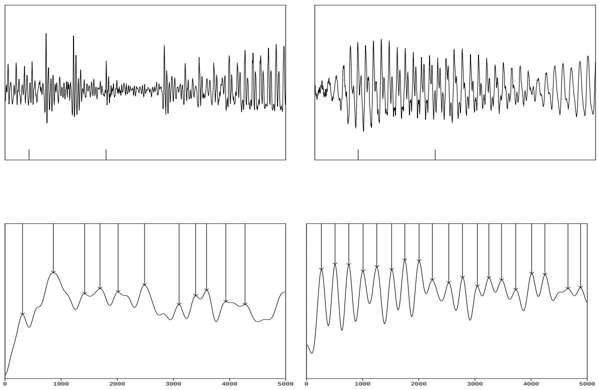

FIGURE 1.2: Speech signal of a laryngealized (left) and a non-laryngealized (right) phone [aI] (top) and the cepstrally smoothed spectra (bottom). The spectra are computed over the analysis windows (38.4 msec) marked by the tics on the x-axis. Significant maxima are marked by arrows. On the left *"war ein"* [v6aIn] and on the right *"Steine"* [StaIn@] was uttered.

1.2.2 APPROACH II (TIME DOMAIN FEATURES):

Since laryngealizations are voice source phenomena they should be easily detectable in the voice source signal. Up to now algorithms for transforming the speech signal into the source signal have mostly been based on inverse filtering using LPC. We achieved best results with a new inverse filtering technique using artificial neural networks (ANNs). We trained a multi-layer perceptron using a database of speech and voice source signals[4] recorded in parallel. The ANN is able to map speech signals into source signals (cf. bottom of figure 1.1) quite accurately [3]. At the moment we use two methods for classifying the output of such an ANN into the three classes unvoiced, non-laryngealized voiced, and laryngealized voiced frames. The first one computes features from the ANN output signal describing the regularity of the signal structure. We cannot yet report any results for this method. The second method uses the sample values of the ANN filter output as input for another ANN (multi-layer perceptron) which is trained to discriminate between the three classes (one output node per class). Up to now best results were achieved with an ANN with 2 hidden layers with 60 and 20 nodes. Input to the ANN was a 38.4 msec window of the (ANN) inversely filtered speech signal, which

[4] Voice source signals (figure 1.1) were measured with a laryngograph .

itself is sampled at 2000 Hz thus resulting in 75 input nodes. The network was trained using the Quickprop algorithm. 65% of the laryngealizations were recognized (12% false alarms).

1.3 Concluding Remarks

In [1] other possible features in the time domain are described. In the near future, we will use these features as well and try to optimize our set of features using e.g. linear discriminant analysis for the selection of features. In addition, a combination of time and frequency domain features will be used. Classifying laryngealizations with ANNs yielded a promising result, and we believe that there is a lot of room for improvement. Comparing the two approaches, one has to take into account that the classifier in approach I has additional information: the cepstral coefficients contain already some information since 25% laryngealizations are found only using them; furthermore, the cepstral coefficients reduce the confusion of voiceless and voiced frames.

In the future we will try to improve word recognition by the inclusion of laryngealized phone classes. It remains to be proven that this information is not already contained in models incorporating context across word boundaries and in mixture densities.

1.4 References

[1] A. Batliner, S. Burger, B. Johne, and A. Kießling. MÜSLI: A Classification Scheme For Laryngealizations. In *Proc. ESCA Workshop on prosody*, pages 176–179, Lund, September 1993.

[2] A. Batliner, R. Kompe, A. Kießling, E. Nöth, and H. Niemann. Can you tell apart spontaneous and read speech if you just look at prosody? In A. Rubio, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F. Springer–Verlag, Berlin, Heidelberg, New York, 1994. (to appear).

[3] J. Denzler, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Going back to the source: Inverse filtering of the speech signal with ANNs. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 111–114, Berlin, September 1993.

[4] W. Hess. *Pitch Determination of Speech Signals*, volume 3 of *Springer Series of Information Sciences*. Springer–Verlag, Berlin, Heidelberg, New York, 1983.

[5] D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation*. PhD thesis, Chalmers University, Göteborg/Lund, 1988.

[6] A. Kießling, R. Kompe, E. Nöth, and A. Batliner. Irregularitäten im Sprachsignal — störend oder informativ? In R. Hoffmann, editor, *Elektronische Signalverarbeitung*, volume 8 of *Studientexte zur Sprachkommunikation*, pages 104–108. TU Dresden, 1991.

[7] P.L. Kirk, P. Ladefoged, and J. Ladefoged. Using a Spectrograph for Measurements of Phonation Types in a Natural Language. *Working Papers in Phonetics*, 59:102–113, 1984.

[8] K. Nebel. Spektrale Merkmale zur Detektion von Laryngalisierungen im Sprachsignal. Diplomarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, 1992.