Visualization of Intelligibility Measured by Language-Independent Features

Tino Haderlein[1,2], Catherine Middag[3], Andreas Maier[1], Jean-Pierre Martens[3], Michael Döllinger[2], and Elmar Nöth[1]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstraße 3, 91058 Erlangen, Germany
Tino.Haderlein@cs.fau.de
http://www5.cs.fau.de

[2] Klinikum der Universität Erlangen-Nürnberg, Phoniatrische und pädaudiologische Abteilung, Bohlenplatz 21, 91054 Erlangen, Germany

[3] Universiteit Gent, Vakgroep voor Elektronica en Informatiesystemen (ELIS), Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

Abstract. Automatic intelligibility assessment using automatic speech recognition is usually language-specific. In this study, a language-independent approach based on alignment-free phonological and phonemic features is proposed. It utilizes models that are trained with Flemish speech, and it is applied to assess dysphonic German speakers. In order to visualize the results, two techniques were tested: a plain selection of most relevant features emerging from Ensemble Linear Regression involving feature selection, and a Sammon transform of all the features to a 3-D space. The test data comprised recordings of 73 hoarse persons (48.3 ± 16.8 years) who read the German version of the text "The North Wind and the Sun". The reference evaluation was obtained by five speech therapists and physicians who rated intelligibility according to a 5-point Likert scale. In the 3-D visualization, the different levels of intelligibility were clearly separated. This could be the basis for an objective support for diagnostics in voice and speech rehabilitation.

1 Introduction

Evaluation of voice distortions is still mostly performed perception-based. This, however, is too inconsistent among single raters to establish a standardized and unified classification. Perception experiments are applied to spontaneous speech, read-out standard sentences, or standard texts. In contrast, already used methods of automatic analysis rely mostly on sustained vowels [9]. The advantage of speech recordings, however, is that they contain phonation onsets, variation of F_0, and pauses [14]. Furthermore, they allow us to evaluate speech-related criteria, such as intelligibility [4]. Intelligibility has been identified as one of the most important aspects of voice and speech assessment [1,13,17]. Experimental tools on intelligibility assessment usually employ an automatic speech recognition (ASR) system to compare the patient's utterance with the target text. However, ASR encompasses acoustic models for representing the basic sounds of a language. Therefore, it can only be used to assess speech of that language. Recently, ASR-free methods were developed [2,11]. They

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 547–554, 2014.

embed acoustic models that can track the phonological properties of the utterance as a function of time. Such a tracking does not rely on what was actually said anymore. It has been demonstrated that these ASR-free techniques are able to assess intelligibility of different voice pathology groups even if the members of the group speak a language that was not included in the training of the phonological models [11].

The results of such an automatic evaluation are basically a sequence of numbers. This will be of no help for the technically uneducated medical personnel. Therefore, the goal of our work is to provide a graphical visualization of a small number of features which are extracted from a high number of "technical" features by some automatic dimension reduction method. The basis of the distance measure between different speakers are the phonological and phonemic features in this study. In order to provide 3-D visualization, two approaches were applied. One is to select the three most relevant features that emerge from an Ensemble Linear Regression involving feature selection. The other is to apply Sammon mapping [15] to the full feature vectors. It allows the graphical representation of abstract data, unveiling underlying structures and configurations. This method itself is not new, but it has never been applied to this concrete problem. In earlier studies, it has been successfully used for the visualization of different levels of voice quality, even from different recording conditions [5,7]. The test set of this study is composed of speech recordings of chronically hoarse speakers. The following questions will be addressed: Can different levels of intelligibility of hoarse speakers be visualized by phonological and phonemic features? Are there also phonological and phonemic properties that can visibly separate the two subgroups of functional and organic dysphonia?

Section 2 describes the test data, Sect. 3 gives an overview on the features obtained from the data, and Sect. 4 reviews the dimensionality reduction methods applied for 3-D visualization. The results are presented and discussed in Sect. 5.

2 Test Data and Subjective Evaluation

73 German persons with chronic hoarseness participated in this study. Patients suffering from cancer were excluded. The most common pathologies were grouped into functional (n=45) and organic dysphonia (n=24; see Table 1). Functional dysphonia is usually caused by too few or too much speaking effort due to psychogenic reasons or vocal misuse. Organic dysphonia has its origin in anatomical changes, such as vocal fold polyps, edemas or pareses. The remaining four speakers suffered from chronic laryngitis. Each person read the phonetically balanced text "Der Nordwind und die Sonne" ("The North Wind and the Sun", [6]), which is frequently used in medical speech evaluation in German-speaking countries. It contains 108 words (71 distinct) with 172 syllables. Additional remarks by the speaker were removed manually from the recordings. The data were recorded with a sampling frequency of 16 kHz and 16 bit amplitude resolution using an AKG C 420 microphone (AKG Acoustics, Vienna, Austria).

Five voice professionals estimated the patients' intelligibility on a five-point Likert scale while listening to a play-back of the recordings. They marked one of the grades "very high", "rather high", "medium", "rather low", or "very low", which were converted to integer values from 1 (very high) to 5 for computation. An averaged mark,

expressed as a floating point value, was calculated for each patient as the mean of the single scores. These marks served as ground truth in our experiments. The inter-rater agreement, computed as the mean correlation of one rater against the average of the four others, is given in Table 1.

Table 1. Number of speakers, age statistics, perceptual evaluation results (intelligibility on a 5-point scale), and inter-rater correlation r for the patient groups

	no. of speakers			age				intelligibility scores				
persons	all	men	women	μ	σ	min	max	μ	σ	min	max	r
total group	73	24	49	48.3	16.8	19	85	2.51	1.02	1.00	5.00	0.82
functional	45	13	32	47.1	16.3	20	85	2.27	1.00	1.00	5.00	0.83
organic	24	9	15	52.2	15.6	25	79	3.06	0.91	1.60	4.80	0.75

3 Features Computed from the Speech Data

The pre-processing stage returns a spectro-temporal representation of the acoustic signal. From this representation, speaker features are extracted which constitute a compact characterization of the speech of the tested person. An intelligibility score is finally computed on the basis of these speaker features.

During the pre-processing stage, a stream of Mel-frequency cepstral coefficients (MFCC) is extracted from the recording. For each 25 ms speech frame (frame shift: 10 ms), 12 MFCCs and an energy value are returned. Based on the stream of MFCCs, two text-independent feature extraction method, focusing on phonological and phonemic aspects, have been explored.

Alignment-free Phonological Features (ALF-PLFs): First described in [12], these features follow from a tracking of the temporal evolutions of the individual outputs of an artificial neural network that was trained (see [10,12] for more details) to generate 14 phonological properties per frame. These properties describe:

- vocal source: voicing
- manner of articulation: silence, consonant-nasality, vowel-nasality, turbulence (referring to fricative and plosive sounds)
- place of consonant articulation: labial, labio-dental, alveolar, velar, glottal, palatal
- vowel features: vowel height, vowel place, vowel rounding

Every phonological property is analyzed by two sub-networks. One of them determines whether the property is relevant at a given time (e.g. it is not relevant to investigate vowel place during utterance of a consonant); the other one determines whether the characteristic (e.g. "labial") is actually present or not.

The hypothesis is that temporal fluctuations in the network outputs can reveal articulatory deficiencies, regardless of the exact phonetic content of the text that was read, at least as long as this text is sufficiently rich in phonetic content. The temporal

analysis of each network output generates a set of parameters, such as the mean and standard deviation, the percentage of the time the output is high (above 0.66), intermediate or low (below 0.33), respectively, the mean height of the peaks (maxima), and the mean time it takes to make a transition from low to high. The overall number of output features is currently 504, and it is acknowledged that several of them may carry similar information. These speaker features are computed without knowledge of the text that was read. Hence, we expect them to be text-independent.

Alignment-free phonemic features (ALF-PMFs): The features, introduced in [10], are based on the hypothesis that intelligibility degradation is correlated with problems in realizing a certain *combination* of phonological classes that is needed for the production of a certain phone. Therefore, the ALF-PMFs follow from a plain analysis of posterior phone probabilities which are themselves retrieved from the phonological properties by means of a neural network. Considering all frames for which the maximal posterior probability is assigned to a particular phone, one computes the mean and standard deviation of that probability, the mean of the peaks (maxima) and the valleys (minima) found in the temporal evolution of that probability. In addition, the percentage of the time a frame is assigned to the phone, and the mean probability of this phone over all frames are computed. Clearly, these features are computed without any knowledge of the text that was read and can therefore be expected to be text-independent. The number of ALF-PMFs is equal to 495.

4 Reduction of Dimension for Visualization

The aim of the dimensionality reduction is to construct a 3-D space that can be visualized using 3-D graphics.

The first method consists of creating a lot of linear regression models on different randomly chosen subsets of the training data and to allow each model to select three features (principle of Ensemble Linear Regression). The three features that were selected most frequently are then retained for visualization.

The second method is Sammon mapping. It performs a non-linear transformation preserving data topology which is represented by a matrix of inter-utterance distances. The distance metric can be chosen without any mathematical restrictions, such as linearity etc. This is the great advantage of the Sammon transform against other dimension reduction operations, such as PCA or LDA [8]. In the pilot experiments, the Euclidean distance between the feature vectors of two respective speakers was used as the distance metric. In the future, we plan to extend the feature set by other feature types which call for another distance metric, e.g. a metric that is more suitable for comparing distances between HMM parameters, as in [5]. The heart of Sammon's method is its special error function E, yielding a stress factor between the actual configuration of n points in an m-dimensional target domain and the original data in a d-dimensional space ($m < d$):

$$E = \frac{1}{\sum_{p=1}^{n-1} \sum_{q=p+1}^{n} \delta_{pq}} \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} \frac{(\delta_{pq} - \nu_{pq})^2}{\delta_{pq}} \tag{1}$$

δ_{pq} denotes the Euclidean distance between feature vectors with number p and q, ν_{pq} is the distance between points $s(p)$ and $s(q)$ in the Sammon map. E is within [0,1], where $E = 0$ means a lossless projection from d- to m-dimensional space. Utterances forming clusters in original space will tend to cluster also in destination space. The same holds for utterances being far apart from each other. In order to achieve the final map, we apply standard steepest descent to (1).

5 Results and Discussion

The three most relevant features retained from the set of ALF-PLF and ALF-PMF features for the representation of intelligibility were:

- *turbulence_presence_meanmax* represents the mean of the peaks in the turbulence feature pattern over the file. It is obvious that intelligibility decreases with rising turbulence in the voice.
- *alveolar_relevance_meanmin* represents the value of the valleys (minima) in the relevance of the property *alveolar*. These valleys are supposed to occur where vowels are spoken. In distorted voices, there is not always a clear distinction between vowels and consonants, and the valleys are less deep than in normal voices. Obviously, this lack of distinction degrades intelligibility.
- *duration* is the overall duration of the read-out text. Since all persons read the same text, the duration is inversely proportional to the speaking rate. It is intuitive that slower speakers are more intelligible. However, a lower speaking rate may also point towards higher speaking effort caused by anatomical problems in the articulatory organs. This has been shown for totally laryngectomized persons with a substitute voice after removal of the larynx [3].

For the separation of functional and organic dysphonia, two features were found to be most relevant:

- *consonantnasality_presence_meanmin* describes the minima of the presence of nasality.
- *h_meanneg* is a phonemic feature describing the mean low evidence that a /h/ has been uttered.

Figure 1 shows a visualization of the most relevant features for each speaker. In the upper left figure, the shading of gray denotes the human reference evaluation of intelligibility. For visualization purposes, the *duration* values were normalized (divided by 5,000), and the *turbulence_presence_max* axis shows the value of $(1 - \textit{turbulence\_presence\_max}) * 30$ in order to achieve about the same range for all three dimensions. The axes were rotated, so that the ability of the features to depict different levels of intelligibility was best visible.

On the right-hand side, Fig. 1 also shows a 3-D visualization computed by the Sammon transform using all available ALF-PLF and ALF-PMF features. The normalization of the feature values was done during the computation of the distance matrix. Again, in the upper image the levels of intelligibility are clearly visible. A transform directly into 2-D did not show such a good result, however.

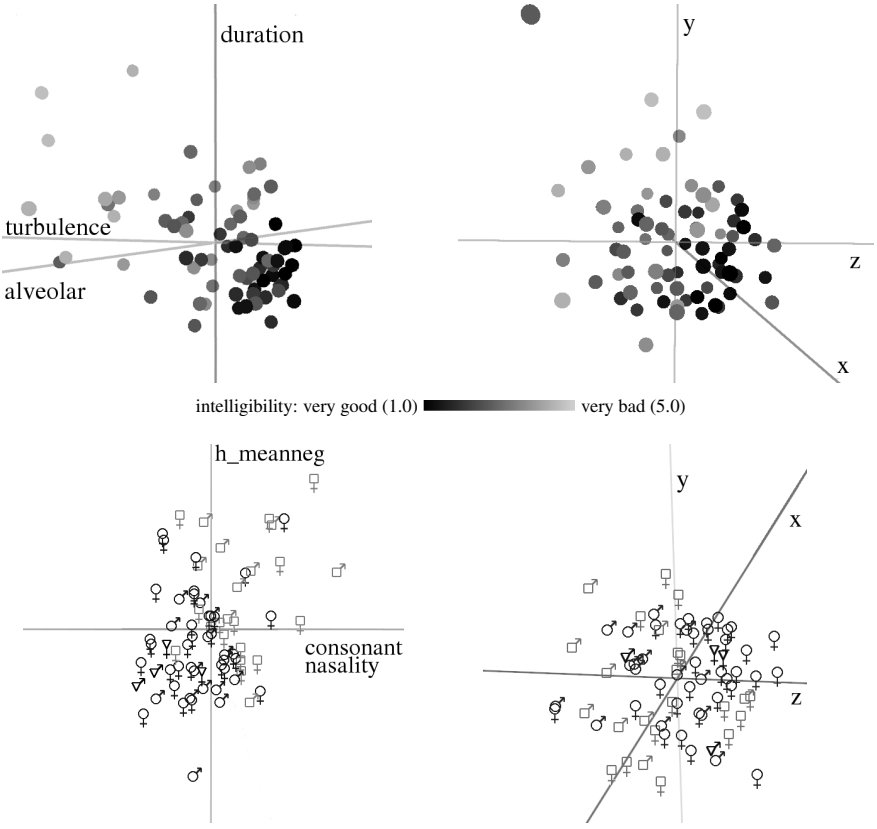

Fig. 1. Visualization of 73 hoarse speakers using the most relevant features only *(left side)* and the Sammon transform on all features *(right side)*; the axes are rotated so that the relevant information is best visible. *Top row:* Human-rated intelligibility is denoted by the shades of gray of the points. *Bottom row:* Functional (○) and organic dysphonia (□), and the four remaining speakers with laryngitis (▽).

In the bottom panels of Fig. 1, the subgroups of hoarseness in the test data are depicted. Both the two most relevant features and all features together resulted in visualizations that show some tendencies for the separation of different diagnoses. The Sammon map shows a tendency to group men to the left and women to the right which is very likely an indicator for F_0. Women with functional hoarseness seem to concentrate in one octant. However, which voice properties are arranged in which direction by the Sammon transform, is dependent on the data and not known in advance. This phenomenon was already reported in [16] where such a map was suggested to have an unlimited number of axes. Most of them represent complex properties of the data and are thus difficult to describe. In the case of the two most relevant features, rather organic dysphonia tends to have higher feature values. The reason could be higher breathiness in these voices. This is confirmed

by preliminary findings by the speech therapists who rated the breathiness on a 4-point scale. For the speakers with functional dysphonia, the average rating was much lower (0.97) than for the persons with organic dysphonia (1.64).

Another research question was the language-independence of the features. In [11], intelligibility models were trained with Flemish and German pathologic speech and tested on the same language, respectively. Here, the test persons spoke German, and the phonological models had been trained with Flemish speech. Additionally, all the test speakers show a similar type of dysphonia, and the training was done with normal speakers. When Support Vector Regression models for intelligibility were trained for the test speakers, the average root mean square error between the computed score and the reference intelligibility was 0.74, corresponding to a Pearson's correlation of r=0.70. This confirms the suitability of the features for evaluation of intelligibility. The basic approach is described in detail in [10].

The results obtained in this study seem to support the following conclusions: Phonologic and phonemic features can be used to display levels of intelligibility. They are even suitable for language-independent analysis. The current feature set can also serve as a basis for the separation of different types of hoarseness automatically. With the integration of more features, the method might be a helpful objective support in the field of voice rehabilitation in the future.

Acknowledgments This work was partially funded by the Else Kröner-Fresenius-Stiftung (Bad Homburg v. d. H., Germany) under grant 2011_A167 and supported by "Kom op tegen Kanker", the campaign of the Vlaamse Liga tegen Kanker VZW and of The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital (Amsterdam). The responsibility for the contents of this study lies with the authors.

References

1. Bellandese, M.H., Lerman, J.W., Gilbert, H.R.: An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. J. Speech Lang. Hear. Res. 44(6), 1315–1320 (2001)
2. Bocklet, T., Riedhammer, K., Nöth, E., Eysholdt, U., Haderlein, T.: Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. J. Voice 26(3), 390–397 (2012)
3. Haderlein, T.: Automatic Evaluation of Tracheoesophageal Substitute Voices. Studien zur Mustererkennung, vol. 25. Logos Verlag, Berlin (2007)
4. Haderlein, T., Moers, C., Möbius, B., Rosanowski, F., Nöth, E.: Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS (LNAI), vol. 6836, pp. 195–202. Springer, Heidelberg (2011)
5. Haderlein, T., Zorn, D., Steidl, S., Nöth, E., Shozakai, M., Schuster, M.: Visualization of Voice Disorders Using the Sammon Transform. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 589–596. Springer, Heidelberg (2006)
6. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)

7. Maier, A., Exner, J., Steidl, S., Batliner, A., Haderlein, T., Nöth, E.: An Extension to the Sammon Mapping for the Robust Visualization of Speaker Dependencies. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 381–388. Springer, Heidelberg (2008)
8. Maier, A., Schuster, M., Eysholdt, U., Haderlein, T., Cincarek, T., Rosanowski, F., Steidl, S., Batliner, A., Wenhardt, S., Nöth, E.: QMOS – A Robust Visualization Method for Speaker Dependencies with Different Microphones. Journal of Pattern Recognition Research 4(1), 32–51 (2009)
9. Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., Corthals, P.: Acoustic measurement of overall voice quality: A meta-analysis. J. Acoust. Soc. Am. 126, 2619–2634 (2009)
10. Middag, C.: Automatic Analysis of Pathological Speech. Ph.D. thesis. Ghent University, Ghent, Belgium (2012)
11. Middag, C., Bocklet, T., Martens, J.-P., Nöth, E.: Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In: Proc. Interspeech, pp. 3005–3008. Int. Speech Comm. Assoc. (2011)
12. Middag, C., Saeys, Y., Martens, J.-P.: Towards an ASR-free objective analysis of pathological speech. In: Proc. Interspeech, pp. 294–297. Int. Speech Comm. Assoc. (2010)
13. Moerman, M.B.J., Pieters, G., Martens, J.P., van der Borgt, M.J., Dejonckere, P.H.: Objective evaluation of the quality of substitution voices. Eur. Arch. Otorhinolaryngol. 261(10), 541–547 (2004)
14. Parsa, V., Jamieson, D.G.: Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J. Speech Lang. Hear. Res. 44, 327–339 (2001)
15. Sammon Jr., J.: A Nonlinear Mapping for Data Structure Analysis. IEEE Trans. on Computers C-18(5), 401–409 (1969)
16. Shozakai, M., Nagino, G.: Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models. In: Proc. ICSLP, Jeju Island, Korea, pp. 717–720 (2004)
17. van As, C.J., Koopmans-van Beinum, F.J., Pols, L.C., Hilgers, F.J.M.: Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. J. Speech Lang. Hear. Res. 46(4), 947–959 (2003)