PERCEPTUALLY INSPIRED FEATURES FOR SPEAKER LIKABILITY CLASSIFICATION

Sira Gonzalez[1,2] and Xavier Anguera[1]*

[1] Telefonica Research, 08019 Barcelona, Spain
[2] Imperial College London, London SW7 2BT, UK

ABSTRACT

We present a novel approach to speaker likability classification. Our algorithm, instead of extracting a large number of features, identifies a small set of features which represent perceptual speech characteristics. For classification, linear support vector machines are used. We train and evaluate the performance on the Interspeech speaker trait challenge database and we show that our likability classifier outperforms the baseline classifier developed for the challenge while considerably reducing the number of features needed.

Index Terms— Speaker traits, likability, classification

1. INTRODUCTION

Speech is the most common mean of communication between humans. The way we are perceived by others when we speak is an important indicator of who we are, and becomes even more relevant when speech is the only signal we receive from our counterpart (e.g. in a telephone conversation). Automatic speaker likability classification can be used to asses when a person's voice will be well or badly perceived by others. Such systems become very useful, for example, as a non-biased indicator for recruitment of telephone assistants or for self evaluation purposes.

The likability of a speaker is a subjective measure and it is difficult to determine objectively what makes a voice agreeable. Rules for breathing, articulation, tone or timing to be a good orator are described in [1]. Moreover, in [2], it was found that attractive voices were associated with lack of tension, presence of confidence and favorable personality ratings. A study to determine the characteristics of dysphonic and normal voices which were important for listeners is explained in [3]. Its findings indicate that naive listeners relied primarily on the fundamental frequency for both voice sets, but also attended to abnormality and breathiness for the pathological voices and to resonance information for normal ones. In [4], they try to determine the suitability of a speaker to serve as a model for building a concatenative text to speech synthesis. They spotted some male-female differences and found a positive correlation of features related to the unvoiced speech power and spectral tilt. The effect of the speaker voice in advertising was studied in [5], where they showed a preference for faster than normal syllable speed and low pitch. In [6], a number of acoustic features were studied and they concluded that temporal features did not have a relationship to the overall rating of a speaker, but features related to the fundamental frequency dynamics and fluency gave significant correlations.

In recent years, a number of speaker likability classifiers have been implemented. In [7], they first focused on the consistency of subjective evaluation of speech likability and found than all sentences from the same speaker were rated similarly but the agreement between different listeners for the same speaker was low. They extracted several parameters and found that the most significant results were mostly dependent on the gender of the speaker. For the classification, around a thousand features used for emotion and affect recognition were extracted, achieving a 69.66% weighted average recall using classification trees on a small database containing 90 speakers. In [8], by also extracting a large number of speech features, specially focused on auditory characteristics, a 67.6% unweighted accuracy was reported also with classification trees on a database later used for the Interspeech 2012 speakers trait challenge. A voice pleasantness classifier using a database containing 77 professional female speakers with radio, theater or other vocal experience is used in [9]. Up to 179 features are extracted from the clinical, quality, intelligibility, naturalness and emotion areas. The final architecture is a support vector machine (SVM) classifier combined with a Gaussian mixture model (GMM)/Bayes methodology in a late fusion scheme. They found out that the best performance was achieved with only 6 features, reporting a classification accuracy of 90.9%.

Recently, the speaker trait challenge proposed at Interspeech 2012 [10] gathered a lot of attention on speaker likability classification by providing a common database and baseline algorithm to work on. In the next section, we will explained this challenge and the different ways in which it was approached. By working on the database provided for the challenge, we propose a new speaker likability classifier focused on understanding the features which make the speech likable. Instead of extracting a large number of features, we extract a small set of meaningful features perceptually inspired and train a linear SVM with them. We evaluate our

*The author performed the work while at Telefonica Research

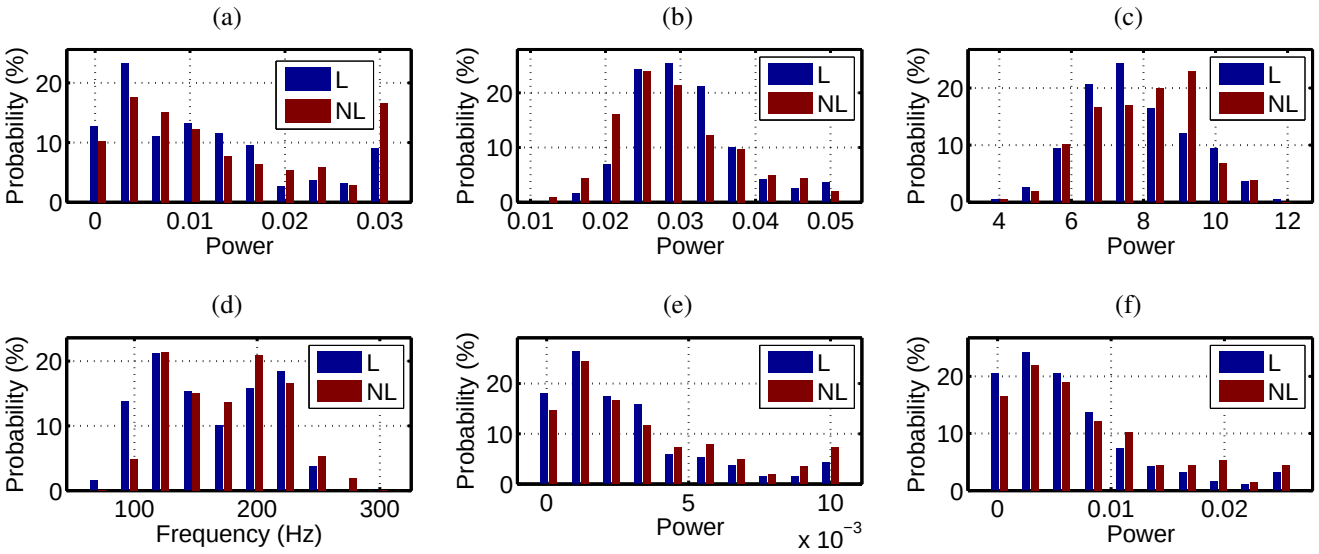

Fig. 1. Histogram for likable (blue) and non-likable (red) voices on the training set [10] for: (a) speech active level, (b) standard deviation of the total power derivative, (c) standard deviation of the power derivative over a 100 ms of the frequency band centered at 2.72 kHz, (d) median pitch, (e) mean speech variance and (f) speech variance standard deviation.

algorithm on the Interspeech speaker trait challenge database and we show that the baseline results can be improved by 3.2% while drastically reducing the number of features from 6125 to 7.

2. INTERSPEECH SPEAKER TRAIT CHALLENGE

The speaker trait challenge proposed at Interspeech 2012 [10] provided a common ground (consisting on a database and a baseline algorithm) for 'perceived' speaker traits: personality, likability and intelligibility of pathologic speakers. The speaker likability database consisted on 800 speakers from a database originally recorded to study automatic age and gender recognition from telephone speech. The features used for the classification, which included several functionals applied to them, were extracted using the openSMILE toolkit [11] and could be classify into three sets: energy, spectral and voicing. The baseline classification accuracy for speaker likability using a total of 6125 features was 55.9% using linear SVM and 59.0% using random forests [10].

Many approaches to the challenge focused on the method used for classification and not so much on the feature set, using the standard features provided by the organizers. Stricted Boltzman machine and deep belief networks [12], anchor models [13] or a new machine learning algorithm based on Gaussian processes [14] were used for the classification. The highest classification accuracy, 64.0%, was achieved by combining the stricted Boltzman machine and deep belief networks.

Among the participants focusing on feature selection, [15] used the Fisher information metric and later on a genetic algorithm. Two methods for feature selection are implemented by [16]: one based on classification and one based on statistical dependence. In addition to all the features from the baseline approach, [17] uses another four kinds of features: basic prosodic features, prosodic polynomial coefficients, spectral features and shifted delta cepstrum features. In [18], they achieved the best performance in the test set, with a classification accuracy of 65.8%, by combining pitch and intonation features with spectral features, using more than $10,000$ features.

In [19], the authors try to understand what makes some voices more likable than others. They subjectively evaluate six characteristics of the speech, finding that likable speakers exhibit almost no perceivable accent, command style or disfluencies. Although the subjectivity of speech likability makes its classification challenging, the low overall performance of the likability classifiers at the challenge may also be related to the short duration of the utterances, which makes the extraction of reliable features difficult.

3. SPEECH FEATURES

As we have seen in the introduction, many approaches extract a large number of features to perform speaker likability classification. Sometimes, the initial set is reduced by feature selection. This makes the number of features falls, but the final selection may not be directly correlated to perceptual characteristics of the speech. In this paper, we focus on the extraction of a small but meaningful set of features which we consider related to the perceptual speaker likability. In some cases, basic functionals are applied to them to extract

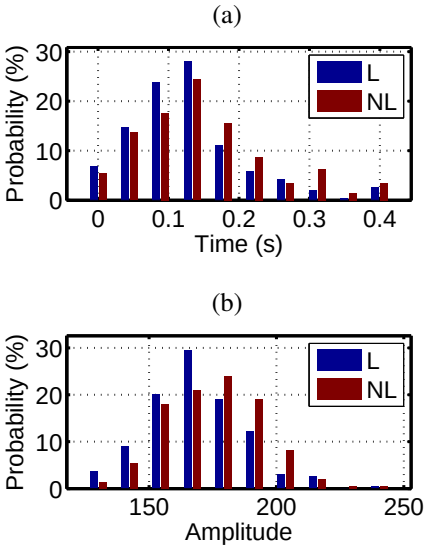

Fig. 2. Histogram for likable (blue) and non-likable (red) voices on the training set [10] for: (a) silent time per second, and (b) mean cross-correlation peak for consecutive frames

the underlying information.

The features are evaluated on the training set of the database to observe their correlation with the speaker likability. If positive results are found, the distribution on the development set is also used to verify this observation.

3.1. Active speech level

Active speech level provides a measure of the loudness of the speech. Under the assumption that all recordings have been made in the same conditions, the different active speech level is only due to the specific speaker.

By evaluating the active speech level of likable and non-likable speakers on the training set of the database using the ITU-T P.56 recommendation [20, 21], we observe a preference for moderate loudness, as seen in Fig. 1(a).

3.2. Variation of the speech power

How the speech power changes in time is provides information about the speaker's accentuation and emphasis. We investigate the total power change with time and the power change in five mel-spaced frequency bands, centered at 261, 621, 1114, 17913 and 2722 Hz. Two approaches are taken into account: power difference in consecutive frames and power derivative using neighboring frames. Frames of 90 ms duration with an inter-frame increment of 10 ms were used.

The most significant results on the training set are obtained for the standard deviation of the total power difference between consecutive frames, Fig. 1(b), and the standard deviation of the power derivative over a 100 ms of the frequency band centered at 2.72 kHz, Fig. 1(c). While in the first case a higher standard deviation is desired, i.e. higher range of variation, in the second case lower standard deviation is more likable. At higher frequencies most power changes are due to fricative sounds, and the lower standard deviation preference may indicate a predilection for smooth fricative transitions or for lower fricative power.

3.3. Fundamental frequency

The fundamental frequency and its functionals have been shown to be an important characteristic in different studies [3, 5, 6]. Therefore, we decide to evaluate its importance in speech likability classification by analyzing the median (for robustness to the fundamental frequency estimator errors), the standard deviation, the mean of the first derivative and the standard deviation of the first derivative. The PEFAC algorithm is used for pitch extraction [22, 21]. We observe that only the median pitch, Fig. 1(d), provides a different distribution between likable and non-likable speakers. Fig. 1(d) illustrates a preference for low fundamental frequencies, related to male speakers. This matches the results showed in [7], which found that higher likability rates are given to male speakers with low fundamental frequency.

3.4. Speech variance

The mean speech variance, a measure of how spread out the speech signal is over a specific period of time, is calculated by segmenting the speech into overlapping frames and measuring the variance in each frame. The frame length was set to 20 ms with 5 ms overlap. We calculate the mean and the standard deviation of this variance over the whole utterance.

The results on the training set show that this feature can provide useful information for the classification, being a lower mean, Fig 1(e), and standard deviation, Fig. 1(f), preferable. This means that it is desirable for the speech amplitude to be uniformly distributed across each time frame.

3.5. Silent segments per second

By measuring the average silent time per second of the speaker we can infer information about how much we should pause when speaking. A speech frame is considered to be silent if its power falls below a specific threshold. Initially, we normalized the speech using the ITU-T P.56 recommendation [20, 21] and we divide the speech into frames of 90 ms duration with an inter-frame increment of 10 ms. The power threshold was set empirically to $7 \cdot 10^{-4}$.

We find that the lower silent time per second is more agreeable, Fig. 2(a). By looking at the spectrogram of likable and non-likable utterances, we observe that the main difference is that while likable speakers tend to keep some power in the word transitions, least likable speakers do not usually have this cadence.

Table 1. Likability classifier features

1	Speech active level
2	Standard deviation of the power derivative
3	Standard deviation of the power derivative over a 100 ms of the frequency band centered at 2.72 kHz
4	Mean speech variance
5	Speech variance standard deviation
6	Silent time per second
7	Mean cross-correlation peak amplitude for consecutive frames

3.6. Correlation between neighboring frames

The smoothness of a speech utterance can be seen as the cross-correlation between neighboring frequencies. We decided to measure the peak of the cross-correlation for frames of 20 ms duration separated from 5 ms to 30 ms.

In Fig. 2(b) we can observe the histogram for likable and non-likable speakers. Opposite to what we were expecting, lower correlation is preferable between consecutive frames. A hypothesis may be that while likable speech has smooth and longer transitions therefore changing its waveform constantly, less likable speech remains longer in a phone (high interframe correlation) and then changes abruptly.

4. LIKABILITY CLASSIFIER

To select the best features for the classifier, we identify the features which provided a good separation on the training set and evaluate that these distributions are maintained in the development set. Table 1 shows the seven selected features. As no difference is made between male and female speakers, the fundamental frequency and its functionals are not taken into account in our algorithm.

The classifier used was a linear SVM with sequential minimal optimization. The training was done on the training set of the database and the soft margin is chosen to obtain the best performance in the development set. The best unweighted average (UA) recall in the development set is 61.6% for a soft margin of 4.92. The accuracy on the training set was 63.2%.

For the results on the test set, we train a linear SVM on the training and development set using the soft margin optimized on the development set. The results obtained for out classifier on the test set of the database are shown in Table 2 together with the baseline performance and the best performance obtained at the Interspeech challenge. We can observe how our classifier is able to increase by 3.2% the baseline performance while reducing the number the number of features from 6125 to 7. The best result of the challenge, 65.8%, was achieved by the fusion of three classifiers each containing a different set of features, some of them in common. Therefore, it is difficult to determine how many features were used to achieve the final classification rate. The lowest limit shown in the table is taken from the individual classifier with the maximum number of features, 10342. Although our algorithm has not reached their performance, the number of features we use is much smaller and we provides not only a likability classification, but also an insight of speech characteristics which could be modified to make the speech more likable.

Table 2. Likability classification results on the test set

	UA(%)	# features	Classifier
Baseline [10]	59.0	6125	Random Forests
Montancie et al. [18]	65.8	> 10000	linear SVM
Proposed	62.2	7	linear SVM

5. CONCLUSIONS

In this paper we have proposed a new speaker likability classifier based on features which represent perceptual characteristics of the speech. We identify a set of features and observe their potential for a speaker likability classification. We then select the best features, a total of seven, and train a linear support vector machine. We show that the baseline performance can be improved by 3.2% while reducing the number of features from 6125 to 7 meaningful features.

6. REFERENCES

[1] J. E. Frobisher, *Acting and oratory: designed for public speakers, teachers, actors, etc.*, New York: College of Oratory and Acting, 1879.

[2] M. Zuckerman, H. Hodgins, and K. Miyake, "The vocal attractiveness stereotype: Replication and elaboration," *Journal of Nonverbal Behavior*, vol. 14, pp. 97–112, 1990, 10.1007/BF01670437.

[3] J. Kreiman, B.R. Gerratt, and K. Precoda, "Listener experience and perception of voice quality," *Journal of Speech, Language and Hearing Research*, vol. 33, no. 1, pp. 103, 1990.

[4] A. Syrdal, A. Conkie, Y. Stylianou, et al., "Exploration of acoustic correlates in speaker selection for concatenative synthesis," in *Proceedings of ICSLP*, 1998, vol. 98, pp. 2743–2746.

[5] A. Chattopadhyay, D. W. Dahl, R. J. B. Ritchie, and K. N. Shahin, "Hearing voices: The impact of announcer speech characteristics on consumer response to

broadcast advertising," *Journal of Consumer Psychology*, vol. 13, no. 3, pp. 198 – 204, 2003.

[6] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations," in *Proc. Interspeech Conf.*, Brisbane, Australia, 2008, pp. 1688–1692.

[7] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proc. Interspeech Conf.*, Makuhari, Japan, Sept. 2010.

[8] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "'Would you buy a car from me?'–on the likability of telephone voices," in *Proc. Interspeech Conf.*, 2011, pp. 1557–1560.

[9] L. Pinto-Coelho, D. Braga, M. Sales-Dias, and C. Garcia-Mateo, "On the development of an automatic voice pleasantness classification and intensity estimation system," *Computer Speech & Language*, 2012.

[10] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, et al., "The interspeech 2012 speaker trait challenge," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[12] R. Bruecker and B. Schuller, "Likability classification - a not so deep neural network approach," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[13] Y. Attabi and P. Dumouchel, "Anchor models and WCCN normalization for speaker trait classification," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[14] D. Lu and F. Sha, "Predicting likability of speakers with Gaussian processes," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[15] D. Wu, "Genetic algorithm based feature selection for speaker trait classification," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[16] J. Pohjalainen, S. Kadiogu, and O. Rasanen, "Feature selection for speaker traits," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[17] M. H. Sanchez, A. Lawson, D. Vergyri, and H. Bratt, "Multi-system fusion of extended conext prosodic and cepstral features for paralinguistic speaker trait classification," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[18] C. Montacie and MJ. Caraty, "Pitch and intonation contribution to speakers's trait classification," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[19] B. Weiss and F. Burkhardt, "Is not bad good enough? aspects of unknown voices' likability," in *Proc. Interspeech Conf.*, Portland, USA, Sept. 2012.

[20] ITU-T, "Objective measurement of active speech level," Mar. 1993.

[21] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997.

[22] Sira Gonzalez and Mike Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.