Classical and Novel Discriminant Features for Affect Recognition from Speech

Raul Fernandez [1,†], *Rosalind W. Picard* [2,‡]

[1]IBM T.J. Watson Research Center. Yorktown Heights, NY, 10598.
[2] MIT Media Lab. Cambridge, MA, 02139.
[†]fernanra@us.ibm.com [‡]picard@media.mit.edu

Abstract

This paper investigates the performance and relevance of a set of acoustic features for the task of automatic recognition of affect from speech using machine learning techniques. Eighty seven novel and classical features related to loudness, intonation, and voice quality, are examined. Using feature selection, the results yield a performance level of 49.4% recognition rate (compared to a human performance rate of 60.4% and a chance level of 20%), while the relevance results show that the more exploratory and novel subset of these features outrank the more classical features in the recognition task.

1. Introduction

In the active research area of recognition of affect from speech it is of particular interest to obtain acoustic features that provide results closer to those of human recognition abilities. While many now "classic" features have been proposed in the literature, their performance has still fallen short of human recognition, suggesting the need to continue a search for novel features and methods. This paper briefly highlights results from an extensive investigation developing new features, and comparing them side-by-side with classical ones using machine learning techniques. (See [1] for many details omitted in this paper.) Algorithms and features associated with modeling loudness, intonation, and voice quality are highlighted in § 2, 3 and 4 respectively, and results of the experiments in §5 with some concluding remarks in § 6.

2. Loudness Features

We seek to investigate a model of loudness which addresses well-known aspects of auditory processing, such as *critical-band* processing, the notion that listening devotes unequal emphasis to different areas of the audible spectrum, and the notion of *masking*, the process whereby the loudness contributed by a sound at a certain frequency is influenced by adjacent artifacts both in time and in frequency.

The features proposed here rely on Zwicker's model of absolute loudness [2], which addresses four stages of loudness processing: (i) attenuating the sound through the outer and middle ear, (ii) filtering by an auditory filter bank to produce excitation levels, (iii) transforming an excitation pattern into specific loudness by a power-law relationship, and (iv) integrating specific loudness across the 24 critical bands of the Bark scale taking into account masking from the excitation levels of adjacent bands. The output of this model consists of a specific loudness pattern (*i.e.*, a distribution over critical bands of specific loudness values) as well as an absolute loudness value obtained by integrating the specific loudness over the Bark scale. The model is suitable for modeling the loudness of a stationary sound. To obtain a time-varying representation of loudness, we extend this procedure by applying a short-time (frame) analysis to windowed segments of speech and building a profile of instantaneous loudness over time.

To introduce a feature set based on this loudness model, let $s_n(n)$ be an RMS-normalized speech signal and K be the number of frames from the short-term analysis of the waveform. Let $N(k)$ and $R(k)$ be the perceptual loudness and RMS value of the kth frame, for $k = 1, \cdots, K$. Let $N'(k,m)$ be the specific loudness pattern for the kth frame, where m covers the Bark scale from $z_1 = 4.25$ to $z_{14} = 17.25$ in increments of 1 Bark. In the following we restrict the analysis to the set of frames $K_{nz} = \{k : N(k) \neq 0\}$ for a which a non-zero perceived loudness is obtained (the model allows for zero loudness if the level of a tone does not exceed what is known as the threshold in quiet, the minimum level that a sound of a given frequency must exceed to be audible.) Let the superscript nz denote a signal evaluated only at frames in K_{nz} and $|K_{nz}|$ be the size of this set. Let the mean perceived loudness be $\mu_N = \frac{1}{|K_{nz}|}\sum_{k\in K_{nz}} N_k$. Let p be the p-th percentile value of the signal $N^{nz}(k)$, and define the p^+-percentile mean $\mu_L^{(p)}$ as the mean of the K_p values exceeding p: $\mu_N^{(p)} = \frac{1}{K_p}\sum_{k:N^{(nz)}(k)>p} N^{(nz)}(k)$. Define, analogously, the p^+-percentile mean $\mu_R^{(p)}$ quantity for $R^{(nz)}(k)$ as $\mu_R^{(p)} = \frac{1}{K_p}\sum_{k:R^{(nz)}(k)>p} R^{(nz)}(k)$. Finally, let the mean specific loudness pattern for the m-th band, delimited by z_m and z_{m+1}, be given by $\mu_{N'_m} = \sum_{k\in K_{nz}} N'(k,m)$. The vector of loudness-related features is defined as follows, letting p be the percentile values 25, 50, and 75:

$$FS_{loud} \doteq [\mu_N, \mu_N^{(25)}, \mu_N^{(50)}, \mu_N^{(75)}, \mu_R^{(25)}, \mu_R^{(50)}, \mu_R^{(75)}, \mu_{N'_1}, \cdots, \mu_{N'_{13}}]^T. \quad (1)$$

The 20-feature vector in (1) is intended to summarize the distribution of the loudness and RMS values over the course of a sound, as well as the specific loudness on different critical bands. Notice that μ_R, the average of RMS values, is not included, as, due to the RMS normalization, this value is expected (and intended) to be less critical. The p^+-percentile features are included, however, since they reflect the distribution of RMS values over time, a feature which may still prove useful.

3. Fundamental Frequency Features

Fundamental frequency has been one of the components of spoken language most studied in the literature on speech and emo-

tion [3, 4, 5]. Since it serves a properly linguistic function, much work has tried to sift out what aspects of F0 are used for encoding linguistic meaning (*e.g.*, contrast between statements and questions) and what aspects serve a paralinguistic function. The latter has been assumed to reside in the continuous variation of parameters (or "gradient features,") such as range and F0 baseline, which can be overlaid on categorical configurations dictated by phonological constraints. Studies suggest, however, that the perception of emotion from F0 does not merely arise from paralinguistic variation, but rather that intonational categories contribute to the perception of affect [4, 5]. Motivated by these observations, we incorporate two types of analyses of F0: a series of statistics summarizing the behavior of the F0 curve (conceptually closer to the gradient features discussed in [4, 6]), and a set of features from the stylized F0 contour.

Let $F0_v(n)$ be the set of voiced values from an $F0$ contour. and consider the following statistics: Let $F0_{v_{skew}}$ be the sample skewness of $F0_v(n)$ and $F0_{>\mu_v}$ be the fraction of voiced values lying above the mean, a measure of how asymmetrical F0 is with respect to its mean. Let iqr_v be the inter-quartile range of $F0$ (a measure of the spread sensitive to outliers), let $range_v^+$ and $range_v^-$ be, respectively, the range of F0 above and below its sample mean (where 95% and 5% percentiles have been used in place of the maximum and minimum values for robustness), and define the following feature vector based on the raw $F0$:

$$FS_{F0-raw} = [skew_v, F0_{>\mu_v}, iqr_v, range_v^+, range_v^-]^T. \quad (2)$$

In [1], we have additionally implemented an F0 stylization based on algorithms presented in [7], which describes the curve in terms of turning points $\{T_p\}$ (defining linear piecewise sections), and which labels every syllable with voicing as pitch accented or unstressed. Let $PA_\%$ be the fraction of pitch-accented syllables and ΔSl_{perc} be the fraction of slope changes in the stylized contour. These measures capture information about the degree to which syllables are accented. Let PA^*_{max} and PA^*_{rg} be the height and range of the strongest pitch accent (over the syllable receiving it). Let m_{decl} be the declination slope of the stylized contour and m_{last} the slope of the last linear segment (a measure considered to reflect boundary tone information), and define the vector of stylized features as

$$FS_{st} = [m_{decl}, PA_\%, \Delta Sl_\%, PA^*_{max}, PA^*_{rg}, m_{last}]^T. \quad (3)$$

A summary of the F0 contour is then obtained from the features derived from the raw F0 representation and its stylization:

$$FS_{F0} = [FS^T_{F0-raw}, \quad FS^T_{F0-st}]^T. \quad (4)$$

4. Voice Source Features

The contribution of voice source to signal para- and extralinguistic information (*e.g.*, attitude, emotion, and vocal pathology) has been investigated and reported in the literature [8, 6]. We next highlight a series of algorithms and features extracted from the glottal excitation which we propose for discriminating between affective categories. We have adopted the linear source-filter model to estimate the glottal volume velocity (GVV) signal by pitch-synchronously inverse-filtering each glottal cycle of the waveform with an all-pole filter estimated during the close-phase of the cycle [1]. Let $\hat{g}_k(n)$ represent this estimate for the kth glottal cycle. The first proposed subset of features is based on a parametrization using the two-phase piecewise Liljencrants-Fant (LF) model of the GVV signal:

$$g(n) = \begin{cases} \frac{-E_e e^{\alpha n} \sin \frac{\pi n}{T_p}}{e^{\alpha N_e} \sin \frac{\pi N_e}{T_p}} & 0 \le n \le N_e \\ \frac{-E_e \ e^{-\beta(n-N_e)} - e^{-\beta(N_0-N_e)}}{1-e^{(-\beta(N_0-N_e))}} & N_e < n \le N_0. \end{cases} \quad (5)$$

Using a constrained optimization algorithm [1], we fit Eq. 5 to each cycle k of the GVV signal to obtain $g_k(n)$ with parameters $\theta_k = [E_{e_k}, T_{p_k}, \alpha_k, \beta_k, N_{e_k}, N_{c_k}, N_{0_k}]^T$. Since the parameter T_p (the point on the open phase at which the GVV reaches zero) depends on the value of the fundamental period or the value of the open phase, we express it as a fraction of one of these constants, and define $\gamma^*_k = \frac{T^*_{p_k}}{N_{e_k}}$, a measure which reflects the asymmetry of the pulse during the open phase. We also let the open quotient $OQ_k = \frac{N_{e_k}}{N_{0_k}}$ be the proportion of the open phase duration to the fundamental period. We now let $S(\{x\}$ be the following set of four statistics on an arbitrary sequence x_k: 25-percentile, median, 75-percentile and inter-quartile range of the normalized difference of x_k, $\quad iqr \ \frac{x_{k+1}-x_k}{x_{k+1}}$, and define a feature vector of LF features as

$$FS^{LF} = \Big[S(E^*_{e_k})^T, \quad S(\gamma^*_k)^T, \quad S(\alpha^*_k)^T, \quad S(\beta^*_k)^T, \cdots \quad S(OQ_k)^T, \quad S(\epsilon^{LF}_{o_k})^T, \quad S(\epsilon^{LF}_{c_k})\Big]^T, \quad (6)$$

where $\epsilon^{LF}_{o_k}$ and $\epsilon^{LF}_{c_k}$ are the sum-of-square errors between the GVV signal and its LF fit for the k cycle through the open and closed phases. This 28-component vector FS^{LF} summarizes statistics of the LF parametrization of the GVV signal over its various cycles, providing a concise description of the distribution of the glottal shape over the waveform. The error parameters have been included to provide a description of the accuracy of the LF fit, and to quantify how the glottal cycles deviate from the idealized LF model. In particular, ϵ^{LF}_o has been included to target phonation types (*e.g.*, breathy or whispery) during which there is some or considerable flow throughout the open phase.

4.1. Other Source Features

In addition to the F0 intonational features discussed, there is a source of F0 variation, this one perceptually associated with voice quality, that is of interest: the short-term period-to-period fluctuations known as *jitter*. Analogous to this kind of F0 variation is a quantity known as *shimmer*, the random short-term changes in the glottal pulse amplitude. Jitter/shimmer measures have been considered in voice quality assessment to describe the kinds of irregularities associated with vocal pathology [9]. Voice pathologies might represent one end of the continuum which these parameters can describe; it is conjectured that between normal and pathological ranges, these parameters may also take on values corresponding to other vocal qualifications.

One approach to quantifying shimmer and jitter, following [10], makes use of the following measures, known as the perturbation factor (PF) and perturbation quotient (PQ_K), defined for any sequence x_n of length N as

$$PF(\{x_n\}) = \frac{1}{N}\sum_{k=1}^{N} \frac{x_{n+1} - x_n}{x_{n+1}} \quad (7)$$

$$PQ_K(\{x_n\}) = \frac{1}{N-K} \sum_{n=(K-1)/2}^{N-((K-1)/2)-1} \frac{x_n - \mu_K(x_n)}{\mu_K(x_n)} \ , \quad (8)$$

where $\mu_K(x_n)$ is a local running average (obtained with a K-point rectangular window). Jitter and shimmer may be quantified by directly applying any of these two measures to, respectively, the time series $\{T_k\}$ of intra-peak intervals (the distance between adjacent instances of maximum excitations) normalized by the sampling frequency, and to E_k, the sequence of intra-peak energy values.

Another proposal to analyze voice quality involves estimating the degree to which the periodic glottal waveform is affected by a noisy component. From the several parameters proposed to model noise in the harmonics, we have chosen the Glottal-to-Noise Excitation Ratio (GNE) since it has been shown to be less colinear with measures of jitter and shimmer, which we are already separately including [11, 1].

The voice quality parameters discussed so far are time-domain features of the speech signal. We consider in addition the Parabolic Spectral Parameter (PSP) proposed by [12] to quantify directly in the frequency domain the decay of the glottal pulse spectrum, a parameter that has been shown to correlate with different phonation types [1].

Based on the features described here, the following 14-D feature vector is obtained, where the perturbation quotients for shimmer and jitter are evaluated over 3, 9, 15, and 21 periods, and the minimum and maximum retained, and $S(\cdot)$ is the set of 4 statistics described earlier:

$$\begin{aligned} FS^{VS} = \; & Jitt_{PF}, \quad Jitt_{PQ}^{min}, \quad Jitt_{PQ}^{max}, \cdots \\ & Shimm_{PF}, \quad Shimm_{PQ}^{min}, \quad Shimm_{PQ}^{max} \cdots \\ & S(\{GNE_k\})^T, \quad S(\{PSP_k\})^{T\ T}. \end{aligned} \tag{9}$$

4.2. Harmonicity Features

This last section presents a novel set of features motivated by psychoacoustics research aimed at quantifying consonance of spectral harmonic patterns [13]. The basic perceptual result relevant to this analysis is a rise-fall curve describing how listeners judge the amount of dissonance/consonance as a function of the interval between two tones [14]. When there is a complex tone, i.e., two sets of frequencies sounding simultaneously, the frequencies interact with each other, depending on their relative strengths and separation, to produce a more complex pattern of rising and falling dissonance. Sethares has proposed a simple linear model, [13], to describe this variation in sensory consonance as a function of frequency interval. The result, known as a dissonance curve, reveals intervals at which a given spectral pattern can sound particularly consonant or dissonant. We derive new "harmonic consonance features" for voiced speech adapted from the Sethares model.

The extension to voice analysis is accomplished as follows [1]. Different voiced sections of speech show different patterns of spectral harmonics, with variations in strength (relative heights) and location (the degree to which the upper partials are multiple of the fundamental). If we remove from the power spectral densities any spurious non-harmonic structure, then the spectral pattern of the idealized harmonics leads to a dissonance curve. Repeating this procedure for near-stationary short-segments of speech then yields a time-varying representation which we call a dissonance diagram. To build a feature set, we first identify a series of landmarks and features on the dissonance curve $D_m(\alpha_n)$ associated with the mth time slice, where n is an index spanning the frequency intervals ($n = 1, \cdots, N$). Let us assume that each interval at which there is maximum dissonance is indexed by α_k^d and every interval of minimum dissonance (excluding unison) by α_j^c. Let us order them such that $D(\alpha_k^d) \geq D(\alpha_{k+1}^d)$ and $D(\alpha_j^c) \leq D(\alpha_{j+1}^c)$, and assume there are K maxima and J minima. Implicit in the dissonance curve is a quantity known as the intrinsic dissonance of the harmonic pattern equal to the sum of the dissonances between all pairs of harmonics within the pattern. Let us denote this quantity as D_{I_m} for the mth time slice. Let us then summarize the landmarks and shape of a curve by the 12-parameter vector:

$$\begin{aligned} D_m^{parm} = \Big[\; & \alpha_1^c, \quad \alpha_2^c, \quad \alpha_1^d, \quad \alpha_2^d, \quad D_m(\alpha_1^c), \quad D_m(\alpha_2^c) \cdots \\ & D_m(\alpha_1^d), \quad D_m(\alpha_2^d), \quad \frac{1}{J}\sum_{j=1}^{J} D_m(\alpha_j^c), \cdots \\ & \frac{1}{K}\sum_{k=1}^{K} D_m(\alpha_k^d), \quad \frac{1}{N}\sum_{n=1}^{N} D_m(\alpha_n), \cdots \\ & \frac{1}{N-1}\sum_{n=1}^{N-1} |D_m(\alpha_{n+1}) - D_m(\alpha_n)| \Big]^T. \end{aligned} \tag{10}$$

This vector contains the locations of the two strongest intervals of dissonance and consonance and their respective values, the average value of dissonance (consonance) peaks (valleys), and the mean value of the dissonance curve and its absolute first difference. Together they summarize major landmarks on the curve, as well as its level and rate of change. For each slice in the dissonance diagram, we collect its intrinsic dissonance D_{I_m} and its parametric summary D_m^{parm}, and define a 14-observation feature vector to summarize the entire dissonance diagram:

$$\begin{aligned} FS^{cons} = \Big[\; & \text{median}_m\{D_{I_m}\}, \quad \text{range}_m\{D_{I_m}\}, \cdots \\ & \text{median}_m\{D_m^{parm}\}^T \Big]^T. \end{aligned} \tag{11}$$

5. Feature Relevance and Applications to Affect Recognition

How well do these features discriminate affective states, and which ones perform best? Since we have $D = 87$ features, exhaustive search of the $2^D - 1$ possible subsets of features is intractable. While only exhaustive methods are optimal, methods such as Sequential Forward Floating Selection (SFFS) [15] have had empirical success in solving feature selection problems. SFFS grows a subset of features by applying a criterion to the left-out features, choosing the feature that optimizes the criterion, and then possibly removing the least significant one at every step. SFFS doesn't guarantee a strictly growing set of features (features may be added to the set, deleted, and then re-added). This is desirable to recover from earlier sub-optimal decisions, but it doesn't lead to a ranking of the features. To explore a feature's relevance to performance, we could account for how often a feature was in the best set of a given size, and weigh those features added earlier over those favored by the algorithm toward the end. We therefore propose the following figure of merit: Let $f_{i,k}$, $1 \leq i, k \leq D$ be an indicator which is 1 if the i^{th} feature was selected when building the best set of size k, and 0 otherwise, and define $w_k = \frac{k-D}{1-D}$ and $FOM_i = \sum_k w_k f_{i,k}$ to be a weighted count of the number of times a feature was selected, where w_k linearly favor features chosen earlier. Evaluating and sorting FOM for all features in the set then yields the ranking. We should note that FOM is just a heuristic to evaluate *individual* relevance since SFFS optimizes feature subsets as *ensembles*.

Rank	Type	Description
1	LF	25-percentile of OQ
2	LF	25-percentile of γ
3	HC	Median of 2nd. highest cons. interval α_2^c
4	HC	Median of average diss.
5	LF	Median value of ϵ_o
6	L	50^+-percentile RMS
7	L	Mean specific loudness (Band 4)
8	LF	25-percentile of ϵ_o
9	L	Mean specific loudness (Band 11)
10	L	Mean specific loudness (Band 10)
11	HC	Median of intrinsic diss. D_I
12	L	Mean specific loudness (Band 9)
13	L	Mean specific loudness (Band 8)
14	L	Mean specific loudness (Band 7)
15	L	Mean specific loudness (Band 6)
16	L	Mean specific loudness (Band 1)
17	HC	Median of cons. values at interval α_1^c
18	L	Mean specific loudness (Band 13)
19	O-VS	Maximum Jitter (Pert. Quotient)
20	L	Mean specific loudness (Band 12)

Table 1: Top 20 features (L=LF Model; HC=Harmonic Consonance; L=Loudness; O-VS=Other Voice Source).

We have investigated the set of proposed features using SFFS with the leave-one-out generalization error of a K-nearest neighbor (K-NN) classifier as a tractable evaluation criterion. Since the features exhibit dynamic structure not directly modeled by a K-NN classifier, we first convert sequences of vectors to a static representation by summarizing them with their sample mean, and then stacking the observations into a single utterance-dependent vector. The data used in this study consisted of a set of speech recordings from 11 actors, delivering a set of 50 sentences in each of the categories *Afraid*, *Angry*, *Sad*, *Happy* and *Neutral*. The results, shown in Table 1, show how the more global features (measured across utterances) describing aspects of loudness and voice quality prove to be the most fruitful in discriminating between affective categories. The loudness features selected include the specific loudness in several bands of the Bark scale, as well as the mean of the integrated perceived loudness across the spectrum. The voice quality features with highest rank include several parameters derived from the LF parametrization of the GVV waveform, as well as several parameters derived from the consonance-based analysis of the spectral harmonics. Other voice source features favored in the selection include jitter measures It is noteworthy that F0 features, often considered important for affect discrimination, are outranked by voice quality and loudness features.

To assess the discriminative ability of this set, Support Vector Machines (SVMs) were trained on the data, and their generalization error estimated through 15-fold cross-validation. Following Principal Component Analysis, SVMs were fit using a Gaussian kernel of width $\sigma = 1, 2.5, 7.5, 10$. The overall recognition rate achieved was 49.4%, which compares favorably with the 60.8% rate of human listeners (and a chance performance level of 20%). The human recognition figure was estimated through a forced-choice listening experiment in which listeners labeled a randomized subset of the data, chosen to be semantically ambiguous in lexical affective content (2 speakers were evaluated by 2 listeners; the remainder by 1 listener).

6. Conclusions

We have explored the ability of a variety of speech features to discriminate affective categories. The feature set includes some classical features (like various statistics derived from the pitch contour) as well as less explored features (like those derived from perceptual models of loudness and harmonic consonance). Our experiments show that the features provide a level of discrimination which, although below human performance, proves competitive. Particularly noteworthy is the result that the more exploratory features outrank some of the more established features in the literature, suggesting a potentially fruitful direction for future research in the area of affect recognition from speech.

7. References

[1] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, MIT, 2003, http://www.media.mit.edu/˜galt/phdthesis.pdf.

[2] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, 2nd ed., ser. Springer Series in Information Sciences. Berlin: Springer-Verlag, 1999, vol. 22.

[3] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *JASA*, vol. 52, no. 4, pp. 1238–1250, 1972.

[4] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[5] S. Mozziconacci, "The expression of emotion considered in the framework of an intonational model," in *Proc. ISCA Wrksp. Speech and Emotion*, N. Ireland, 2000, pp. 45–52.

[6] K. R. Scherer, D. R. Ladd, and K. E. A. Silverman, "Vocal cues to speaker affect: Testing two models," *JASA*, vol. 76, no. 5, pp. 1346–1356, November 1984.

[7] P. C. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Ph.D. dissertation, The University of Edinburgh, 1994.

[8] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Comm.*, vol. 40, no. 1-2, pp. 189–212, April 2003.

[9] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *JASA*, vol. 103, no. 3, pp. 1628–1639, March 1998.

[10] H. Kasuya, Y. Endo, and S. Saliu, "Novel acoustic measurements of jitter and shimmer characteristics from pathological voice," in *Proceedings Eurospeech 1993*, K. Fellbaum, Ed., vol. 3, Berlin, 1993, pp. 1973–1976.

[11] D. Michaelis, T. Grams, and H. Strube, "Glottal-to-noise excitation ratio – a new measure for describing pathological voices," *ACUSTICA*, vol. 83, pp. 700–706, 1997.

[12] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter - A new method for quantification of the glottal flow," *Speech Communication*, vol. 22, pp. 67–79, 1997.

[13] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. London: Springer-Verlag, 1998.

[14] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *JASA*, vol. 38, pp. 548–560, 1965.

[15] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. on PAMI*, vol. 19, no. 2, pp. 153–158, 1997.