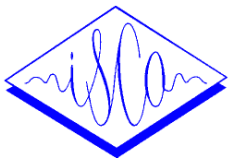

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

PROSODIC PARAMETER FOR SPEAKER IDENTIFICATION

Katarina Bartkova, David Le Gac, Delphine Charlet, Denis Jouvet

France Télécom R&D – DIH/IPS
2 av. Pierre Marzin, 22300 Lannion, France
Katarina.Bartkova@rd.francetelecom.com

ABSTRACT

This study investigates to what extend prosodic parameters are speaker dependent and, therefore, can be used in a speaker identification system. Fundamental frequency, phone duration and phone energy are investigated and modeled under different forms and their efficiency to identify successfully the speaker, to whom the model belongs, is evaluated. The speech material used to model and evaluate the prosodic parameters is collected from 28 speakers and consists of free spontaneous speech. For 61% of the speakers one or several very efficient prosodic cues are found yielding an average ranking lower, thus better, than the 3rd best position. For 36% of the speakers the average ranking of the right speaker is between the 3rd and the 5th position. Only one speaker out of 28 is not satisfactorily represented by any of the prosodic parameters.

1. INTRODUCTION

Text-independent speaker recognition has large potential application field (speaker segmentation of vocal archive, speaker identification…). At present, in this research field, gaussian mixture modeling (GMM) constitutes the state of the art [12]. However, such a modeling does not account neither for temporal aspect of the speech signal nor its prosodic organization. Furthermore, the acoustic modeling is sensitive to the communication channel and to the surrounding noise. It appears, that speakers are characterized by prosodic parameters (speech rate, phoneme and pause duration, end of utterance lengthening…) that are less affected by the quality of the communication channel. Moreover, prosodic parameters are highly complementary with the HMM acoustic modeling. The aim of the present study consists in investigating speaker specific prosodic parameters. These parameters are to be extracted and processed automatically by a speaker identification system. The main benefit expected from this approach consists in increasing system robustness to the communication channel.

Studies dedicated to speaker characterization focus mostly on spectral properties of the speech signal. It has been shown [10] that the segmental domain contains a great amount of speaker specificity. However, little is known about inter-speaker variation of prosodic parameters, such as fundamental frequency, intensity and temporal organization of the utterance.

In [4] it is suggested that prosodic aspects of speech contribute probably more to speaker characterization than the speaker segmental realization, for the range of the prosodic variability is large and more difficult to structure than segmental units. In an extensive study of idiosyncrasy in prosody, [7] showed that though prosodic cues are not sufficient to identify a speaker, however, they can highly contribute to it.

Fundamental frequency (F_0) is the most frequently investigated prosodic parameter. Specifically, the average value of F_0, its standard deviation and distribution are often described as the most efficient cues for speaker identification ([5], [6], [7], [8], [9], [10], [11] & [14]). Furthermore, F_0 contour comparison and their alignment with the segmental level are also helpful in speaker characterization ([1], [7] & [13]). These studies, however, focus on text reading or specific speech styles such as, for instance, sport comments ([7]).

Many authors have investigated temporal aspects of utterances such as articulation rate, pauses and segmental durations. It is claimed that articulation rate (number of syllables per second) can be a reasonably good cue to characterize speakers ([7], [8] & [9]), however, it is not as speaker-specific as F_0 is. [7] and [8] point out that pauses (silent and/or filled) as a proportion of the speech signal duration can characterize a speaker. However, this parameter varies considerably for the same speaker in different speaking conditions, and as for segmental duration, a speaker-specific use is found in different reading styles. In [4] a large individual variation is observed in segmental duration distribution between stressed and unstressed syllables, but again this speaker specification is not consistent from one reading style to the other. Studies on French syllable durations in interaction with speech rate and syllable position in sentence [2] show that mainly in utterance-internal position vowel duration is an efficient cue in speaker characterization.

Little attention is paid to energy in phonetic studies. However, in different speech styles intensity seems to contribute to speaker identification [7].

The goal of this study is to find prosodic parameters that can be used in an automatic text-independent speaker recognition system. Thus, there are two main constraints. First, the prosodic parameters are to be extracted automatically. As it is a text-independent mode, no phonetic knowledge of the speech utterance is available. When it is required, phoneme segmentation is obtained by using an unconstrained phonetic HMM-based decoding. Such decoding leads to a certain amount of errors, hence, the prosodic parameters investigated must be robust to segmentation errors. The second important constraint is the amount of training and testing data per speaker. Depending on the application, the amount of training and testing data can be very small and parameters models are to be estimated correctly with only a few seconds of speech.

After a presentation of the experimental framework (section2), the prosodic parameters are described (section 3) and their evaluation is presented and discussed (section 4).

2. EXPERIMENTAL FRAMEWORK

In this paper the various prosodic parameters are studied individually. In order to evaluate their efficiency, the parameters, chosen according to phonetic knowledge, undergo modeling. For each parameter a model is thus built for each speaker. Then these models are used to identify the speaker in the test procedure.

For the purpose of this study, simple models are used. Most of the time, they are based either on discrete densities, which are defined by the frequency of occurrence of the given parameter in different value intervals (histograms), or on gaussian densities, which are defined by the mean value and standard deviation of the corresponding parameters. When discrete densities are used, the speaker identification relies on computing distances between the speaker models (histograms) and the histogram computed on the test utterance. When gaussian densities are used, the speaker identification relies on the cumulated log-likelihood of the parameters extracted on the test utterance, likelihood computed on each speaker model.

The database used in this work was originally designed at France Telecom R&D to study keyword indexing for voice mail retrieval. It contains 28 speakers recording up to 40 telephone messages each. The style is spontaneous speech and the average duration of the messages is about 10s. The database was recorded over several weeks. 5 utterances of each speaker are used to estimate the speaker model for the prosodic parameters under study. The remaining utterances (from 5 to 35 per speaker) are used to evaluate the efficiency of the parameters.

3. PROSODIC PARAMETERS

The parameters investigated are the three main prosodic parameters: fundamental frequency, phone duration and energy. For each defined parameter, the modeling used is described.

3.1. Fundamental frequency

Fundamental frequency is probably the best of the three main prosodic cues as it contains good speaker specificity not only between male and female speakers, but also between speakers of the same sex having lower or higher average F_0 values. Moreover, F_0 is more independent of the communication channel than energy or spectral coefficients as it is not or only slightly affected by the noise as well as by the channel specificities. Unlike phoneme duration, F_0 is independent of any possible HMM segmentation errors, for segmentation into phonetic units is not absolutely required. Several attempts for modeling fundamental frequency are investigated.

Phone F_0 and **Frame F_0:** Discrete models are estimated using the absolute F_0 value in Hz for each voiced frame (frame F_0) or each voiced phoneme (phone F_0). When phone F_0 is modeled, one value per voiced phoneme is used, which corresponds to the average F_0 values over the frames associated to this phoneme. Frame F_0 modeling is segmentation-free and every voiced frame F_0 value is used. These discrete models capture not only the speaker's average F_0, but also the extreme values of his F_0 scale. The extreme values reflect the speaker's habits while the distribution around his mean F_0 value reflects his intrinsic characteristics.

Frame F_0 slope: Some speakers have larger F_0 movement variations than others. Therefore it seems reasonable to investigate F_0 slope variations. The slope modeling captures the direction and the amplitude of the F_0 movement. The slope is obtained by estimating the derivative of the F_0 on an 11 frame-window. Histograms show a clear difference between speakers characterized by a large scale of F_0 movements and those that have a more flat way of speaking (see Figure 1).

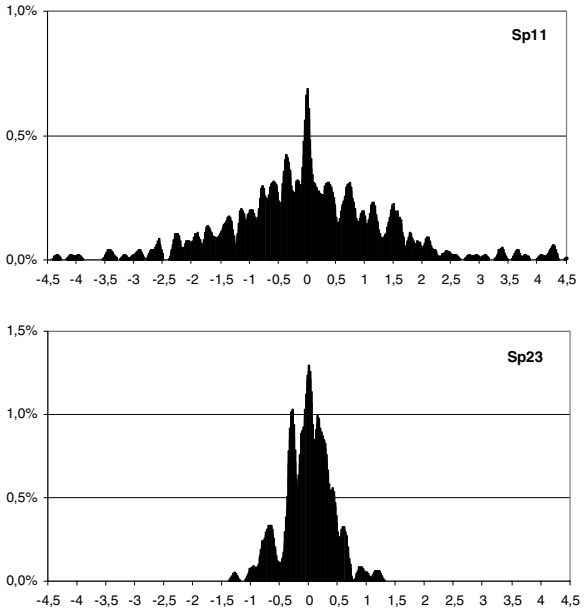

Figure 1: Frame F_0 slope smoothed histograms for 2 speakers.

Prosodic unit: One of the particularities of French spontaneous speech is a rising movement of the F_0 on the major prosodic boundaries. In fact, the rising F_0 pattern constitutes here more than 90% of all the F_0 patterns observed in utterance non-final position. The amplitude of this movement can be speaker dependent: some speakers exaggerate this movement attaining an extra high F_0 target, while others remain more moderate. To model F_0 values on prosodic group boundaries, the speech signal is segmented automatically into prosodic units. The prosodic boundary decision is based on the amplitude of the F_0 movement compared to the speaker average F_0 value and standard variation on every part of the signal between two pauses. A visual inspection of some data shows that the boundary detection is reliable and no major boundaries are omitted. The discrete modeling is applied on F_0 values extracted from the last (rising or falling) movement of the prosodic group.

Prosodic pattern: An attempt is made to model the F_0 pattern all over a prosodic unit. This way, the F_0 movement is modeled together with its temporal evolution. The speech signal is first segmented into prosodic units. Each rising or falling movements of F_0 are captured by three values (two extreme and one intermediate), hence the pattern length depends on the number of F_0 movements inside the prosodic unit. Though the F_0 pattern varies according to the utterance content, it reflects also the speaker prosodic rhythm and habit. Therefore it can be expected that the average distance to his own patterns (as found in the training data) will be smaller than the average distances to the other speakers' patterns. A test pattern is compared only to patterns of similar length, and patterns are temporally aligned

Speaker	Phone F_0	Frame F_0	Frame F_0 slope	Prosodic unit	Prosodic pattern	Speech rate	Phone duration	Phone energy	**Symbolic notation**
Sp01	**2.4**	2.7	2.7	3.7	2.7	8.9	13.9	7.3	+++
Sp02	**1.4**	2.0	**1.4**	2.2	1.6	1.8	8.0	10.8	++++
Sp03	6.5	5.3	14.6	6.3	7.5	12.9	14.2	**3.9**	++
Sp04	**1.9**	2.1	5.8	2.5	4.1	10.9	4.8	13.5	++++
Sp05	9.3	10.3	**4.8**	6.8	14.4	14.6	12.3	27.1	+
Sp06	4.1	5.9	12.6	**3.2**	3.6	12.9	4.2	11.1	++
Sp07	5.0	6.7	9.5	6.1	13.0	21.5	**3.1**	8.7	++
Sp08	7.4	9.7	21.3	**7.1**	10.6	19.4	9.2	12.1	---
Sp09	4.3	4.0	24.5	5.2	2.8	12.0	19.7	**3.4**	++
Sp10	6.7	5.9	10.2	6.1	7.8	6.9	22.3	**4.0**	+
Sp11	1.5	1.4	9.3	1.3	**1.1**	9.4	13.5	5.8	++++
Sp12	**1.5**	**1.5**	9.0	1.8	**1.5**	17.7	3.1	7.4	++++
Sp13	4.4	3.6	24.9	4.5	**2.7**	14.7	2.8	21.1	+++
Sp14	3.8	3.7	3.5	3.1	10.5	10.3	**1.0**	14.3	++++
Sp15	3.3	**1.3**	10.9	2.2	2.2	10.7	15.0	8.6	++++
Sp16	**4.2**	6.3	16.5	6.4	11.0	8.6	8.1	9.1	+
Sp17	**4.6**	6.0	22.2	7.2	6.4	9.4	12.0	26.3	+
Sp18	5.2	5.2	16.7	4.9	10.6	13.1	**2.2**	15.8	+++
Sp19	**1.9**	**1.9**	16.1	2.2	3.0	9.6	12.2	13.9	++++
Sp20	3.3	4.0	**1.7**	3.0	6.5	12.9	10.9	8.1	++++
Sp21	3.9	**2.1**	7.7	3.7	3.4	13.7	18.4	5.1	+++
Sp22	**2.3**	2.7	21.5	3.0	2.9	8.5	20.0	17.8	+++
Sp23	2.8	2.0	28.0	**1.5**	1.6	8.1	12.8	14.8	++++
Sp24	5.6	6.1	6.2	4.6	10.2	7.8	15.6	**3.3**	++
Sp25	5.1	4.6	10.6	6.9	**4.1**	5.5	10.5	15.1	+
Sp26	1.5	**1.2**	15.8	2.6	1.7	7.6	2.0	18.0	++++
Sp27	4.6	2.7	24.1	4.6	1.8	10.2	6.7	**1.1**	++++
Sp28	**1.4**	1.6	2.0	2.3	1.7	8.6	19.1	2.0	++++

Table 1: Average true-speaker ranking for the different prosodic parameters and a global symbolic notation.

for handling the timing differences, before computing a distance between them.

3.2. Phone Duration and speech rate

Phone duration distribution in the sentence is speaker dependent. The difficulty of its use in text independent speaker recognition is due to the HMM segmentation errors. In fact, when an HMM system is used to segment a speech chain without lexical and syntactic constrains, phoneme deletions and insertions are frequent. Such a phonetically poor segmentation can corrupt the use of phoneme duration in automatic processing. Moreover, in a non-constrained segmentation no hypothesis on the word end can be emitted. However, the content word last syllable is stressed in French and is therefore lengthened. Some speakers are easily detectable by an exaggerating last syllable lengthening in stressed position. Unfortunately hypothesis on stressed position when the sentence content is unknown is prone to error. Another drawback of text-free speaker recognition is filled pause occurrences in spontaneous speech. A filled pause can be an independent vowel with a flat F_0 but also the last syllable of any (content or function) word with a non-flat F_0. Therefore filled pauses are not easy to detect and eliminate automatically and can also corrupt the speaker speech rate calculation and its use.

Speech rate is modeled by the mean vowel duration and its standard deviation. For speaker identification, a score is calculated vowel-by-vowel by comparing its duration to the speech rate gaussian model of each speaker. These scores are then averaged to determine the speaker scores for identification.

Phone duration: A discrete phone duration modeling is carried out for each speaker. The discrete modeling captures the phoneme duration distribution all over the sentence. The tail of the histogram can be considered as speaker dependent. It contains the possible speaker characteristic lengthening of the stressed syllables. The lengthening of stressed syllables by some speakers turned out to be a consistent cue.

3.3. Energy

Speech energy is less independent from the communication channel and the surrounding environment than the other two prosodic parameters. Nevertheless, an attempt is made here to investigate to what extend this parameter can be used. The energy modeling approach is motivated by the fact, that some speaking styles are based on syllable energy distribution. In French the stress, which is normally situated on the last syllable of the prosodic unit, can move, under some condition, to the first syllable of a content word. Such a shift, perceived as didactic stress, is very common in journalistic style for instance. But this stress shift can also be speaker-dependent. Hence, a discrete modeling of the normalized vowel energy value is carried out. Each vowel energy value is normalized by the highest vowel energy value over the utterance.

4. PARAMETERS EVALUATION

As mentioned before, in order to evaluate the parameters and their complementarities, speaker identification tests are carried out using each prosodic parameter individually. No attempt is made in this study to combine the prosodic parameters together, or with the cepstral coefficient used by the state of the art GMM-based speaker recognition systems.

For each parameter its performance is measured as the average ranking at which the true speaker model identifies the test speaker. The best ranking is 1, which means that the speaker is correctly identified by his own model. The worst ranking corresponds to the total number of the speakers, which means that the speaker's own parameter model provides the worst result on the test utterance. Closer the average ranking value is to 1 and better the speaker is characterized by this parameter. These average rankings are reported in table 1 for the various prosodic parameters and speakers. For each speaker the best ranking is highlighted and the efficiency of the corresponding prosodic parameter is symbolically indicated in the last column, from excellent ("++++") to very poor ("---").

Complementarities of the parameters tested can be determined. The same parameter is not necessarily to be used for all the speakers. Each speaker can have his best parameter or set of parameters by which he is characterized. It is also observed, that, when discrete modeling is carried out, focusing on a specific part of the model (i.e. using only part of the histogram) can provide better results on some speakers than using the whole histogram. This is the case for instance for F_0-slope. Some speakers use more falling F_0 patterns than others. Therefore, the left side of the F_0-slope histogram represents them more accurately than the right side, which stores rising slopes. Further investigation is to be made on the modeling.

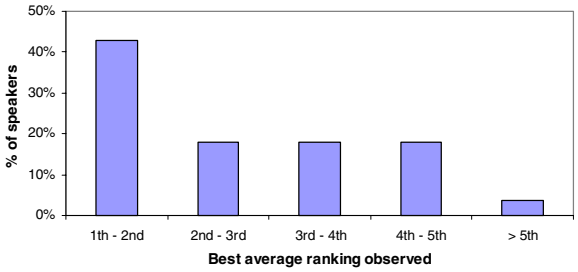

Figure 2: Speakers' best average ranking.

Figure 2 illustrates the percentage of speakers' best average rankings. For example, the first bar indicates that 43% of the speakers have obtained a best average ranking between 1 and 2 with one of the prosodic parameters. 81% of the speakers have obtained a best average ranking between 1 and 4.

The best results are obtained by phone F_0 modeling followed by frame F_0 modeling. The speech rate turned out to be the less efficient parameter of all those studied in this paper. The reason probably stems from the fact that speakers' speech rate is not constant and it can vary considerably from one sentence to the other. Such acceleration or slowing down of the speech rate yields a large standard deviation value which makes the gaussian modeling ineffective. This is also due to an unreliable estimation of the model parameters because of the limited amount of training data available.

5. CONCLUSION

It is shown in the present study that prosodic parameters such as fundamental frequency, phone duration and phone energy can efficiently characterize a speaker. This study focuses on the phonetic aspect of the prosodic parameter determination. Using phonetic knowledge, prosodic parameters are automatically extracted from the speech signal and modeled either by gaussian densities or by discrete densities. Then they are evaluated individually on spontaneous speech. Another experiment, conducted on an independent data set collected from different speakers, has provided similar results to those presented here-before.

Prosodic parameters can be used as complementary information by an automatic system in speaker identification tasks. The advantage of these parameters is that some of them are not, or only very slightly, corrupted by the communication channel quality whereas cepstral coefficients, widely used for speaker recognition, are more affected by noise or channel distortions. For that reason it is reasonable to suppose that these parameters can improve the performance of the speaker identification systems currently in use.

No decision is taken here on how to use prosodic parameters in a speaker identification system or what kind of modeling is the most appropriate. The models described and tested in this paper are quite simple ones thus one can reasonably expect that a more appropriate parameter modeling will improve over the results presented here.

REFERENCES

[1] Atal B.S., "Automatic speaker recognition based on pitch contours", *in JASA* 52:1687-1697, 1972.

[2] Duez D., "How articulation rate and position in utterance and phrase affect segmental duration : within- and cross-subject variability", *in RLA2C*, Avignon, 16-19, 1998.

[3] Fant G., Kruckenberg A., Nord L., "Rythmical structures in text reading. A language contrasting study", *Eurospeech 89,* Paris, 1:498-501, 1989.

[4] Fant G., Kruckenberg A., Nord L., "Prosodic and segmental speaker variations", *Speech Communication* 10, 521-531, 1991.

[5] Jiang M., "Fundamental frequency vector for a speaker identification system". *Forensic Linguistics* 3, 95-106, 1996.

[6] Kraayeveld J., Rietvelt A.C.M. & van Heuven V.J., "Speaker Specificity in Prosodic Parameters", in *Working Papers*, Dept of Linguistics and Phonetics, Lund 41, 1992.

[7] Kraayeveld J., *Idiosyncrasy in Prosody*, PhD dissertation, Nijmegen, 1997.

[8] Künzel H.J., "Some general phonetic and forensic aspects of speaking tempo", *Forensic Linguistics*, 4/I:49-83, 1997.

[9] Magrin-Chagnolleau I., Imitation des voix: étude de paramètres prosodiques et spectraux, mémoire DEA, Université Paris 3, 1996.

[10] Nolan F., *The Phonetic bases of speaker recognition*, Cambridge University Press, Cambridge, 1983.

[11] Nolan F., "Speaker Recognition and Forensic Phonetics", in W. Hardcastle and J. Laver (eds), *A Handbook of Phonetic Science*, Blackwell, Oxford, pp. 744-767, 1997.

[12] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Models", in *Speech Communication*, 1995, vol.17, n.1, pp.91-108.

[13] Vaissière J., "Caractérisations des variations individuelles du contour de fréquence du fondamental observées dans des phrases lues en anglais", *JEP*, Trégastel, 87-92, 1994.

[14] Van Dommelen W., "The contribution of speech rhythm and pitch to speaker recognition", *Language and Speech* 30(4):325-338, 1987.