INVESTIGATIONS INTO PROSODIC SYLLABLE CONTOUR FEATURES FOR SPEAKER RECOGNITION

Marcel Kockmann[1 2], Lukáš Burget[1] and Jan "Honza" Černocký[1]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2]SVOX Deutschland GmbH, Munich, Germany
{kockmann|burget|cernocky}@fit.vutbr.cz

ABSTRACT

We investigate various ways of generating prosodic syllable contour features that have recently been applied to enhance systems for speaker recognition. We compare different approaches for segmentation of speech into syllable-like units, techniques for contour modeling and the extraction of pitch and energy, taking into account the computational complexity and gender dependence. We show that the performance is especially affected by the segmentation and the quality of the pitch tracking algorithm and that the features are highly gender dependent. Still, computationally simple ways of segmentation of speech can be used to achieve good results, as experiments on 2006 NIST speaker recognition evaluation task indicate.

Index Terms**—** Speaker recognition, prosodic features, syllable contours

1. INTRODUCTION

Recent National Institute of Standards and Technology (NIST) evaluation for speaker verification systems has shown that the use of prosodic information to enhance acoustic state-of-the-art systems has become very popular [1, 2, 3, 4]. While most participants use classical prosodic features like duration, energy and pitch in a long temporal context, the actual realizations diverge.

The way of segmenting speech into units which are suitable for prosodic modeling differs a lot. Although most use a syllable-like context, the segmentation techniques span from simple energy-based decisions [2] up to accurate syllables derived from a speech recognition system [1]. Furthermore, there is a huge disparity in the way of modeling speech segments. Besides extracting characteristics like the mean, maximum or minimum [1], continuous feature trajectories are often modeled by Gaussian Mixture Models (GMM) [2, 3, 4]. Various approaches for curve fitting, as well as different ways of handling undefined pitch values appear. Finally, the algorithms for extracting the basic prosodic features differ a lot.

This work takes into account the need for a deeper investigation into the creation of continuous syllable contour features which are suitable for Joint Factor Analysis modeling (JFA) [5], and proved their capability to enhance state-of-the-art acoustic systems [2]. Several techniques for segmentation, contour modeling and basic feature extraction are compared and experimentally evaluated, and we show huge dispersion in accuracy and computational requirements. The performances of the proposed systems are presented in terms of equal-error-rate (EER) for the text-independent NIST SRE 2006 speaker identification task [6].

The organization of the paper is as follows: Section 2 describes different configurations for extracting syllable based features, including basic prosodic features themselves, the way the utterance is segmented into syllable-like units and the actual modeling of the temporal trajectory of the basic features. Section 3 presents the experiments and results, and conclusions are given in section 4.

2. FEATURE GENERATION

Generating prosodic contour features mainly consists of three parts:

1. Extraction of basic prosodic features.
2. Segmentation of speech into syllable-like units.
3. Approximation of temporal feature trajectories.

2.1. Segmentation

Various approaches for segmentation of speech into long-temporal units are presented, starting from the most computationally complex to the simplest.

LVCSR syllables: Syllables are created from the word output of a Large-Vocabulary-Continuous-Speech-Recognition (LVCSR) system[1] using human-created rules [7]. The phone alignments of the recognized words are used to generate correct English syllables. The example in Figure 1 depicts how the recognized word *weird*, with its phones *w+ih+r+d*, leads to a single syllable. This way of segmenting is highly language-dependent but the segmentation is accurate. Compared to other methods, segments are relatively long.

Pseudo syllables: Pseudo syllables are generated from the output of a phone recognizer as described in [8]. Each vowel serves as nucleus and surrounding consonants as onset and coda. Our phonetically rich Hungarian phone recognizer [9] is taken as a language-independent detector. Segments are also relatively long. In the example in Figure 1 the resulting segment is nearly the same as for the LVCSR.

Phone boundaries: The boundaries from the phone recognizer are directly taken as the segments, which results in more shorter segments. It is doubtful if they are that suitable for prosodic modeling (see also Figure 1), as they rather represent a trend like falling or rising, than the character of the whole contour.

Vowel Onset Points: Syllable segments are determined by Vowel

This work was partly supported by European project MOBIO (FP7-214324), by Grant Agency of Czech Republic project No. 102/08/0707 and by Czech Ministry of Education project No. MSM0021630528.

[1] Many thanks for providing their SNERF features to Luciana Ferrer and SRI International.

Onset Points (VOP) as presented in [10]. The strength of excitation shows a significant change at the transition from consonant to vowel. Source excitation information is approximated by the Hilbert envelope of the linear prediction residual. The area between two detected VOPs is taken as a syllable, which often results in very long segments, especially in speech pauses. In Figure 1 it can also be observed that this method splits connected segments rather inappropriately.

Energy valleys: As proposed by [5], the normalized energy is directly used to segment speech. As indicated in Figure 1, the local minima of the energy contour determine the segments, with a minimum length of 60ms. This technique also often results in short segments, due to fluctuations in the energy signal.

Fixed window: Like for standard acoustic features, we propose a fixed overlapping window, but with a long temporal context of 300ms and a shift of 150ms. For the example in Figure 1, this span covers nearly the whole word *weird*, but of course the segmentation is quite arbitrary, depending on the frame shift.

2.2. Contour modeling

This subsection describes the methods used to actually model the temporal trajectory of pitch and energy as well as treatment of undefined pitch values in speech segments.

2.2.1. *Curve fitting algorithm*

As proposed in [8], the trajectory can be modeled by a Discrete Cosine Transform (DCT). Taking the n leading coefficients gives de-correlated features independent of the segment length. Another approach is presented in [5]. The feature segment is modeled by taking the coefficients of an n-th order Legendre polynomial. Both methods translate characteristics of the curve, like mean, slope and finer details into the feature vector. The grade of the detail can be controlled by using more/less DCT coefficients or in-/decreasing the polynomial order, respectively. In preliminary experiments, best results could be achieved with six coefficients.

2.2.2. *Voiced/Unvoiced*

A specific problem in modeling continuous pitch is that this feature is undefined in unvoiced regions. The simplest approach is to use only frames with valid pitch values and collapse these frames before modeling the contour. With Legendre polynomials, it is also possible to keep the undefined values as gaps and to model the contour from the first valid pitch value to the last. Another possibility is interpolation of pitch to close the holes. For energy, one can either take the same frames as for pitch or the whole segment.

2.3. Basic prosodic features

We investigate three popular pitch tracking algorithms that are implemented in *Snack/Wavesurfer* [11] and *Praat* [12]. *Snack* is used in two different modes, as they give significantly different results: *ESPS* uses the normalized cross-correlation function with dynamic programming and *AMDF* stands for Average Magnitude Difference Function. The algorithm implemented in *Praat* is also based on autocorrelation (we will simply call it *Praat*).

Figure 1 illustrates their general behavior on a simple example. The *ESPS* mode gives much smoother values while *AMDF* produces rough steps in the pitch contour, due to quantization effect, where the step-size is proportional to the fundamental frequency. Also, we observe that *AMDF* tends to produce more halving/doubling. The *Praat* output is similar to *ESPS*, but produces generally less pitch values. Note that *Snack* is used with its default settings while we configure *Praat* as in [5].

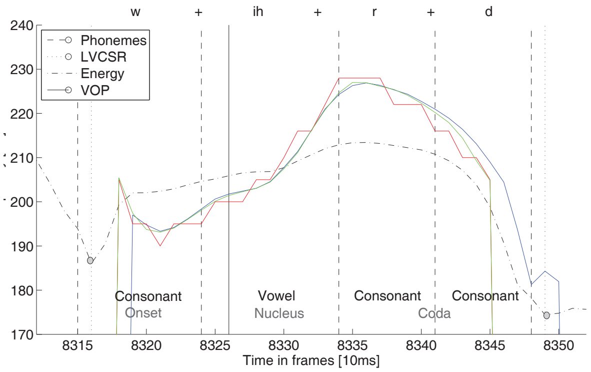

Fig. 1. Example for different segmentations and pitch extractions for the word *weird*. ESPS (blue), AMDF (red) and Praat (green).

As the range of pitch values is gender dependent, we will also investigate the effect of normalizing pitch values and the influence on a gender independent modeling approach.

Furthermore, two energy extraction methods are compared. As is common use for acoustic features, we compare the usage of normalized energy as in [5] and approximation of energy by the 0th cepstral coefficient.

3. EXPERIMENTS

3.1. Setup

For all experiments 13 dimensional feature vectors are generated which comprise segment length (in 10ms frames), and 6 coefficients for pitch and energy each.

As a back-end, we use a GMM-JFA framework [13] as described in [3], which uses low dimensional subspaces to model speaker- (eigenvoices $\mathbf{V}$) and intersession-variability (eigenchannels $\mathbf{U}$). Prior to estimating the subspaces, gender-dependent Universal Background Models (UBMs) with 128 Gaussians are obtained by Expectation-Maximization (EM) Training. Discrete as well as continuous features are used within one feature vector, so variance flooring is crucial while EM training. Variances are floored to $1/100$ of the global variance.

Our standard configuration uses 50 eigenvoices and 20 eigenchannels per gender. We initialize matrices $\mathbf{V}$ and $\mathbf{U}$ by PCA [14] and iteratively retrain first $\mathbf{V}$ and then $\mathbf{U}$. Contrary to [3], we do not use the residual matrix $\mathbf{D}$, so the speaker is only modeled by the 50 speaker factors. For small amounts of test data, the integrative scoring over the channel distribution [15], rather than using point estimate of the channel, has proved to be beneficial. On this task, we have only several hundred frames per test utterance (depending on the type of segmentation), compared to several thousands for acoustic short term features. Finally, all scores are normalized with zt-norm [16] in a gender-dependent way.

Table 1. *Pitch: ESPS, Energy: C0, Modeling: DCT, voiced only. Different segmentations.*

Segmentation	EER [%]
Fixed window	12.1
Energy valleys	13.7
Vowel Onset Points	15.8
Phone boundaries	13.2
Pseudo syllables	12.5
LVCSR syllables	**11.2**

3.2. Data

Experiments are performed on the core condition of the NIST 2006 Speaker Recognition Evaluation (SRE) [6], which contains English trials only. The 1-side training 1-side test condition is considered, where approximately 2.5min of speech is available from a 5min telephone conversation to train each speaker and for each test trial. This set contains 329 female and 248 male training utterances (where multiple utterances can arise from one speaker) and 23687 test trials. Results are presented in terms of equal-error-rate[2]. The UBMs as well as the eigenvoice- and eigenchannel subspaces are trained on all-English one-conversation utterances from the NIST 2004 and 2005 SRE data sets. 300 z-norm utterances and 100 t-norm models per gender are taken from NIST 2004 database.

3.3. Results

The first experiments were carried out to compare different segmentation techniques. *Snack ESPS* pitch and C0 energy were modeled with six DCT coefficients using only the voiced frames. As shown in Table 1, the type of segmentation affects the EER about 30% relative. It is interesting to see, that the complexity of the segmentation mostly corresponds to the results. The most accurate LVCSR syllables give the best rate with 11.2%, while the energy performs nearly the worst. With a huge degradation, compared to all other segmentation methods, the VOPs seem quite unsuitable, probably because syllable end points are not detected properly. So often, connected pitch contours over syllables are cut at the VOP (like indicated in Figure 1) and merged with another fragment over speech pauses. Surprisingly, the most simple way of fixed windows results in the second best error-rate of 12.1%. Capturing long context seem to be important, as all methods that generally result in shorter segments perform worse. The results of the fixed-frame segmentation may indicate, that long time span is even more crucial than correct phonetic alignment of the syllable-like units.

The following experiments show the effect of different contour modeling and further consolidate segmentation results. The setup is kept, only the curve fitting algorithm is switched from DCT to Legendre polynomials, as described in Section 2.2.1. Results in Table 2 show the same trend, best EER for LVCSR with 11.4%, nearly worst for energy with 14.1%, while DCT modeling generally leads to little lower error-rates. The advantage of both methods, compared to, for instance, a simple polynomial curve fitting is, that they operate on orthogonal basis functions and result in de-correlated features, necessary for GMMs with diagonal covariances. In preliminary experiments we observed, that even additional de-correlation with Principal Component Analysis (PCA) could not lead to same performance for simple polynomial curve fitting.

Table 2. *Pitch: ESPS, Energy: C0, Modeling: Polynomials, voiced only. Different segmentations.*

Segmentation	EER [%]
Fixed window	12.1
Energy valleys	14.1
Vowel Onset Points	17.5
Phone boundaries	13.6
Pseudo syllables	12.6
LVCSR syllables	**11.4**

Table 3. *Pitch: ESPS, Energy: C0, Modeling: Polynomials. Different treatment of unvoiced regions.*

Treatment of unvoiced	EER [%]
Voiced frames only for f0 and energy	11.4
Voiced f0 range, keep gaps, same frames energy	11.1
Voiced f0 range, keep gaps, all energy	**11.0**
Interpolation of f0, all frames f0 & energy	11.7

In addition to the curve fitting algorithm itself, processing of undefined values is explored with LVCSR segmentation setup from Table 2. We compare four ways:

1. Using only voiced frames for pitch and energy.
2. Using pitch from first to last voiced frame in the detected segment, but keeping possible holes in the pitch trajectory and using the same frames for energy.
3. The same frames for pitch as in 2., but using all energy in the segment.
4. Linear interpolation of pitch, using all frames for pitch and energy.

In Table 3, generally better results are achieved when the contour is modeled over the gaps, which suggests that preserving the pitch trajectory structure is important. Best result of 11% is achieved with third method, so even use of energy in unvoiced regions enhances the modeling. Interpolation of pitch in unvoiced regions seems to harm rather than help, mainly due to many segments that will result in a straight line for pitch.

Furthermore, the influence of the basic prosodic feature generation is evaluated experimentally, with setup used for Table 1, but changing pitch and energy extraction methods. Table 4 indicates that the quality of pitch estimation highly affects the overall EER. While *Praat* and *ESPS* perform equally, the "steps" and general quality of *AMDF* contour seem to harm a lot and EER drops to 13.6%.

When comparing the two energy extraction methods experimentally, using normalized log-energy instead of C0 approximation also decreases the performance (see also Table 4) to 11.8%.

Finally, we apply utterance-based mean-normalization of pitch values prior to curve fitting, with gender-dependent (same setup as for Table 1) and gender-independent configuration (same setup, but identical UBM and JFA-subspaces for male and female), respectively, to investigate the influence of these features on a gender independent system. Feature vectors only differ in the first coefficient for pitch, which represents the mean of the segment. While we get about 10% relatively better performance with a non-normalized feature set (compared to normalized features) on a gender-dependent system, the non-normalized features perform approximately 10% relatively worse when used with a single gender-independent setup. Results

[2] Note that evaluation key det3 version 9 from NIST was used to measure the system performance.

Table 4. *Modeling: DCT, voiced only. Different pitch and energy extractions.*

Pitch	Energy	EER [%]
Snack ESPS	**C0**	**11.2**
Snack AMDF		13.6
Praat		11.3
Snack ESPS	E	11.8
	C0	**11.2**

Table 5. *Pitch: ESPS, Energy: C0, Modeling: DCT, voiced only. Effect of Normalization and gender dependence.*

Normalized	Gender dependent	EER [%]
No	**No**	14.8
Yes		**13.6**
No	**Yes**	**11.2**
Yes		12.5

in Table 5 point out that the features are highly gender-dependent and that normalization of pitch is crucial to build a single gender-independent model. Still, when used in a gender-dependent setup, the un-normalized pitch contours represent the speaker characteristics better.

4. CONCLUSIONS

We have evaluated many different techniques for the creation of prosodic syllable contour features. It is shown that the quality of segmentation into syllable-like units mostly corresponds to the achieved error rate. As the computational complexity and the language constraints also increase, the proposed fixed-length temporal windows bear a computationally inexpensive alternative with only 8% relative degradation in performance, compared to an accurate segmentation to syllables. Generally, capturing of long temporal units seem to be very important, probably more than a correct linguistic segmentation.

For the contour modeling itself, both methods are suitable to approximate the temporal trajectories of feature streams. An important attribute is that the algorithm translates the contour to de-correlated coefficients. Slightly better results are obtained with Legendre polynomials when the original pitch structure is preserved with its gaps, instead of collapsing the features. Modeling the whole energy in the speech segment, even where no pitch is detected, further enhances the performance. This suggests, that also the unvoiced parts of the speech signal covers speaker information, that can be employed in a prosodic system (Acoustic systems usually make use of all speech frames, no matter if voiced or unvoiced).

Both implementations of the examined pitch algorithms based on auto-correlation perform equally, while the quality of AMDF algorithm is not that suitable for prosodic modeling. Generally we have observed, that halving/doubling and arbitrary pitch values, produced in unvoiced regions, highly affect the performance, because the approximated curves are corrupted in these segments. This has to be considered when the pitch tracking algorithm is parameterized: Rather less, but more reliable, than many scattered pitch values. For energy features, we have observed that 0th cepstral coefficient outperforms the raw energy, like it is often seen in other speech applications. The Mel-Filter based weighting of the signal energy seems to be more appropriate.

Furthermore, syllable contour features are highly gender-dependent due to different pitch ranges of male and female speakers which must be considered when building the system in a gender-independent way. Still, the un-normalized features, namely the absolute mean of the pitch in the segment, seem to have more discriminative power, when used in a gender dependent setup.

To summarize, the best examined configuration uses a gender-dependent system with pitch from *Snack* in *ESPS* mode, C0 as energy feature, Legendre polynomial approximation of pitch and all energy and an accurate syllable segmentation from an LVCSR system, and results in EER of 11% on NIST SRE 2006 task. Compared to the results of the prosodic sub-system in [2], where fusion of prosodic and high-performing acoustic system resulted in enhancement of over 10% relatively, we obtained slightly better results for the proposed configuration, while there is still potential to improve, as our JFA model is trained on much less data.

5. REFERENCES

[1] Kajarekar, S. et al., “THE SRI NIST 2008 speaker recognition evaluation system” In: Proc. of ICASSP ’09, Taipei, 2009, p. S. 4205-4208

[2] Kenny, P, Dehak N, Ouellet, P, “The CRIM Systems for the NIST 2008 Speaker Recognition Evaluation” In: Proc. 2008 NIST Speaker Recognition Evaluation Workshop, Montreal, CA, NIST, 2008, p. 1-4

[3] Burget, L. et al., “ BUT system description: NIST SRE 2008”, In: Proc. 2008 NIST Speaker Recognition Evaluation Workshop, Montreal, CA, NIST, 2008, p. 1-4

[4] Yan, Y. et al., “Description of IOA Systems for SRE08 - Speaker Recognition Evaluation” In: Proc. 2008 NIST Speaker Recognition Evaluation Workshop, Montreal, CA, NIST, 2008, p. 1-4

[5] Dehak, N. et al., “Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification”, in Audio, Speech, and Language Processing, September 2007, Volume 15. pp. 2095–2103.

[6] “The NIST Year 2006 Speaker Recognition Evaluation Plan”, Online on: http://www.nist.gov/speech/tests/spk/2006.

[7] Shriberg, E. et al., “Modeling prosodic feature sequences for speaker recognition”, in Speech Communication Volume 46, Issues 3-4, July 2005, Pages 455-472

[8] Kockmann, M. and Burget, L. “Contour Modeling of Prosodic and Acoustic Features for Speaker Recognition”, in proc. of Spoken Language Technology, 2008, p. 45-48.

[9] Schwarz, P. et al., “Hierarchical structures of neural networks for phoneme recognition”, in Proc. of ICASSP, Toulouse, 2006.

[10] Mary, L. and Yegnanarayana, B. “Extraction and representation of prosodic features for language and speaker recognition”, in Speech Communication Volume 50, Issue 10, October 2008, Pages 782-796

[11] Sjölander, K., “The Snack Sound Toolkit”, Online on: http://www.speech.kth.se/snack.

[12] Boersma, P., “Praat, a system for doing phonetics by computer”. Glot International, 2001, 5:9/10, 341-345. http://www.praat.org/

[13] Kenny, P. et al., “A Study of Inter-Speaker Variability in Speaker Verification”, in IEEE Trans. Audio, Jul 2008, Vol. 16, p. 980-988.

[14] Burget, L. et al.,“Analysis of feature extraction and channel compensation in GMM speaker recognition system,” in IEEE Trans. on Audio, Speech and Language Processing, September 2007.

[15] Glembek, O. et al. “Comparison of Scoring Methods used in Speaker Recognition With Joint Factor analysis”, in Proc. of ICASSP, Taipei,2009, p. 4057-4060.

[16] Auckenthaler, R. et al., “Score normalization for text-independent speaker verification systems”, in Digital Signal Processing, 10/2000.