interspeech 2008 incorporating SST 08

Foreign accent identification based on prosodic parameters

Marina Piat, Dominique Fohr, Irina Illina

Speech Group, LORIA-CNRS & INRIA
BP 239, 54600 Vandoeuvre-les-Nancy, France
"http://parole.loria.fr/", email: { piat,fohr,illina }@loria.fr

Abstract

In this paper we propose an automatic approach for foreign accent identification. Knowledge of the speaker's origin allows to adapt acoustic models for non-native speech recognition. In this study, we use a statistical approach based on prosodic parameters. This approach relies on the fact that prosody is different between languages, and has been done within the framework of the HIWIRE (*Human Input that Works In Real Environments*) European project. The corpus is composed of English sentences pronounced by French, Italian, Greek and Spanish speakers. Results obtained with duration and energy are promising for foreign accent identification: 67.1% correct L1 identification with duration and 68.6% with energy. These two parameters combined with MFCC achieve a 87.1% correct foreign accent identification rate.

Index Terms: non-native speech, foreign accent identification, prosodic parameters

10.21437/Interspeech.2008-235

1. Introduction

The efficiency of an Automatic Speech Recognition (ASR) system deteriorated significantly when confronted with non-native speech. To improve ASR performance, models should be adapted to the speaker's origin. Thus, speaker's origin has first to be detected and identified.

Some studies have suggested that the speaker's first language (L1) can be identified using statistical methods based on MFCC.

Different models have been designed. A global model (GMM) was used for Chinese accent identification: a GMM model was trained for each accent and each gender using MFCC parameterization. Arslan and Hansen [1] identified Turkish, German, Chinese and English accents in American speech with word HMM models for each foreign accent. Kumpf and W.King [2] have discussed the identification of three foreign accents in Australian English speech using HMM (Hidden Markov Models) phone models for each accent. Bouselmi et al. [3] proposed the use of discriminative phoneme sequences to detect speaker's origin for four accents (French, Italian, Greek and Spanish) in English speech.

All these works used cepstral parameters but other parameters can be employed, in particular prosodic parameters. Several works have also addressed the perception of prosodic features of foreign languages. Horgues [4] showed that prosody plays a significant role in the perception of French accents in English speech. Ramus [5] concluded that syllabic rhythm was a necessary and sufficient cue for French adult subjects to discriminate English accents from Japanese sentences.

Some works used prosodic features for foreign accent identification. Arslan and Hansen [6] mainly used fundamental frequency, energy and the four first formants as features for monophone HMM models to identify one accent among four (English, Turkish, Chinese and German) in American speech. Kumpf and King [7] obtained speaker's accent classification rates close to the benchmarks set by human listeners: the speaker accent classification rate reached 84.7% for automatic accent identification versus 87.4% for human listeners on 6 seconds speech segments. The feature set included MFCC coefficients, phoneme segment duration, F0, delta F0 and contextual information. In the context of Arabic dialect identification, the study [8] of long-term prosodic variations and short-term micro-variations permitted an identification for three "Arabic dialects".

Previous research works have shown that prosodic features are relevant cues for accent identification because prosody is different for languages. The goal of our paper is to design an automatic system for foreign accent identification which includes prosodic features. Our proposed approach is based on the following prosodic parameters: duration, energy and fundamental frequency. Compared to [7] we studied each prosodic parameter independently to assess the contribution of each parameter. Compared to [6], where the authors used isolated speech, our work deals with continuous speech to identify foreign accent.

Section 2 of this paper presents the context of this study: the HIWIRE European project. The choice of prosody at syllabic level is explained in section 3. Section 4 presents our modelization. Experimental results are detailed in section 5.

2. Context

This work has been conducted within in the framework of the HIWIRE project.

2.1. The HIWIRE project

The HIWIRE (*Human Input that Works In Real Environments*) project (2004-2007) was a European project involving four research laboratories in Spain, Italy, France and Greece, along with two companies *Loquendo* in Italy and *Thalès* in France. One goal of this project was to recognize pilot speech in an airplane cockpit. Pilots have to speak English, but are not always native speakers.

2.2. The HIWIRE corpus

The recorded corpus within the HIWIRE framework is composed of utterances corresponding to aeronautic task pronounced by 70 Europeans: 20 French, 20 Greek, 20 Italian and 10 Spanish. Each speaker pronounced about 100 utterances composed of a few English words. For example:``Next'',``Range forty'',``Cannot accept nine four four''
The lexicon was composed of 134 application words. Sampling

Accepted after peer review of full paper

frequency was 16 kHz.

3. Motivations and methodology

The use of acoustic models adapted to the speaker's origin improves non-native speech recognition. The aim of our work is to identify the first language of non-native speakers from prosodic features (F0 temporal variation, syllable duration, and intensity).

Our speech corpus is made up of word sequences, which cannot be understood outside the domain of aeronautics, except for some short sentences, such as "change frequency". These word sequences are expected to be pronounced with the intonation pattern of a declarative sentence without special emphasis on a given word (no narrow focus accent). So, in our corpus we have to deal for the most part with lexical accent.

Lexical accents fall on a given syllable of a word. For a given language, the place of the lexical accent is either fixed, i.e. the accent always falls on the same syllable of the word (e.g. the first word syllable in Czech, the last in French), or free, i.e. the place of the lexical accent varies with words (e.g. English).

Some speakers tend to keep their native characteristics when they speak a foreign language. In our application, when a non-native person speaks English he or she may make accent errors related to his or her mother language. This is especially true for subjects who have a very poor grasp of the foreign language.

In this paper, our methodology is based on HMM modeling prosodic parameters for each speaker group with the same L1.

The syllable bearing the lexical accent is often characterized by a lengthening of syllable duration, an increase in energy and a variation in F0. According to the speaker's L1, lexical accent would be more or less marked. That is why we modelized variations in duration, energy and fundamental frequency at the syllable level for each word in the corpus. The aim was not to find the place of the accent in the word but to take it into account automatically.

First, we study each prosodic parameter independently. For that purpose, each parameter is used as feature vector for stochastic models. For the sake of comparison, we also exploited MFCC parameters which are known as robust ASR parameters. Then, a combination of the best prosodic parameters with MFCC coefficients was evaluated.

4. Modeling

Three features have been developed and evaluated: duration, energy and F0. Each feature is modeled by an HMM. One HMM model is built for each foreign accent and each syllable. For instance the HMM models for the syllable "*z-ih*" of the word "*zero*" pronounced by a French speaker is noted "*F_z-ih_zero*".

4.1. Duration

In some languages, the lexical accent is mainly marked by a lengthening of the accented syllable of the word. When a person speaks a foreign language, he or she tends to keep his native habits. As has already been mentioned, the place of the lexical accent within the word varies with the language. As an example, French speakers would tend to lengthen the last syllable of foreign words. Using syllable duration differences could help in identifying the speaker's L1. Due to variations in the rate which a speaker speaks (some subjects speak more slowly than others), the absolute duration of a syllable is not relevant. This is why we study the ratio of syllable duration over word duration. This ratio is modeled by a 1-state HMM.

4.2. Energy

Energy varies from one syllable to another depending on the place of the lexical accent. The syllable bearing the lexical accent will often have strong energy. A non-native speaker may not mark it well. So, for the same word, the distribution of energy could vary according to the speaker's L1. Energy also varies from one speaker to another (according to voice intensity). For this reason energy must be normalized. To do that, we subtract the maximum energy value (dB) on each utterance. Energy is extracted at each frame from the signal, then normalized. We also used first and second energy derivatives.

We try two methods for energy modeling:

- In the first method, the energy evolution of each syllable is divided into three parts (beginning, middle, end) and modeled by a 3-state HMM.
- A second method consists in computing the average energy of each syllable. This average energy is modeled by a 1-state HMM.

4.3. Fundamental frequency

In most languages lexical accents are also marked by fundamental frequency variations.

F0 also varies from one speaker to another, so we need to normalize its values too. F0 is extracted at each frame of the signal using the Snorri software [9].

Several methods have been tried using F0. Only the most basic method and the one giving the best results are presented:

- A first method consists in using values normalized with a warping coefficient specific to each speaker (corresponding to the vocal tract length [10]) and then in studying the evolution of these values during the syllable (3-state HMM).
- A second method consists in computing the slope of F0 during the syllable (1-state HMM) in order to model the global emphasis of the syllable of the word (this method will be referred to "F0 slope" for the rest of this article). The slope is computed from 3 values (value at the previous, current and next frame) using the Least Square Root method.

4.4. MFCC

We compared our results to baseline results based on MFCC. Energy was not included in MFCC. A 25ms Hamming window was applied followed by a 24 filter bank for extracting the 12 MFCC coefficients and their first and second derivatives. 3-state syllable HMM models for each word and for each speaker's L1 were trained using MFCC features.

4.5. Foreign accent identification system

Experiments were performed using cross validation: a group of 7 speakers of all origins was chosen as a test group, and the rest used for training (63 speakers). This process was repeated until all the speakers are tested.

Monosyllabic words and acronyms were discarded from the vocabulary (example of an acronym: *VHF*). Due to the small

Table 1: *Percentage of correctly identified speaker's origins*

Parameter	Vector dimension	Identification rate (%)
Normalized F0	1	28.6
F0 slope	1	54.3
Average energy	3	60.0
Duration	1	67.1
Energy	3	68.6
MFCC	36	82.9

size of our corpus, only the most frequent words were retained for evaluation. The set V of selected words is: *zero*, *request*, *seven*, *select*, *mayday*, *weather*, *accept*, *level*, *performance*.

For each studied parameter, syllable HMM models were trained for each word and each accent. 16 Gaussians by HMM state are estimated.

The evaluation phase was performed as follow: the test speaker pronounced a few sentences which transcription was supposed to be known. The L1 identification was performed in two steps. During the first step, an automatic segmentation of the sentences was performed: a forced alignment was done to obtain the beginning and the end of every word. Secondly, for the words belonging to the vocabulary V, syllable recognition guided by the transcription of the word into syllables was performed: syllable models of the four accents were used. So, we obtained a succession of syllables of different origins O_i. The origin $\hat{i}$ which gathers the largest number of syllables of origin O_i was affected to the speaker. For example if we know that the test speaker pronounced "*zero seven*" and the sequence of HMM models recognized are: `F_z-ih_zero I_r-ow_zero I_s-eh_seven I_v-ah-n_seven`, then the speaker would be identified as an Italian.

5. Experiments and results

5.1. Results for prosodic parameters

For each prosodic parameter, Table 1 gives the percentage of speakers whose origin is correctly identified. (i.e. the identification rate). Using only the normalized F0 values of the syllable, the percentage of correct accent identification is low (28% of the speakers). This is perhaps due to artifact values at the border between voiced and unvoiced segments. A better modeling for the fundamental frequency is the slope of the syllable (54.3%). The computation of the slope is less influenced by abnormal values.

60% of origins are identified from the average energy computed per syllable, while energy identifies 68% of origins. Duration would also appear to be a good indicator since it permits recognition of origin for 67.1% of speakers. F0 variation obtains the worst results in our study. One explanation could be that poor speakers tend to flatten their melodic curve.

To compare overall results, MFCC coefficients obtain the best identification rate: 82.9% of the speaker's L1 are correctly identified. However, we can notice that MFCC feature vectors have 36 components, while prosodic parameters have only 1 or 3.

Table 2, 3, 4 and 5 present confusion matrices (in %) of correctly identified speaker's origins (French (F), Italian (I), Greek (G) and Spanish (S)).

In Table 2, French, Italian and Greek speakers are very well identified (90%-100%), while Spanish are not (10%). These

Table 2: *Confusion matrix in percentage of identified speaker's origins using **MFCC** (average 82.9%)*

	F	I	G	S
F	**100%**	0%	0%	0%
I	5%	**90%**	5%	0%
G	0%	5%	**95%**	0%
S	20%	30%	40%	**10%**

Table 3: *Confusion matrix in percentage of identified speaker's origins using **energy** (average 68.6%)*

	F	I	G	S
F	**65%**	10%	5%	20%
I	25%	**65%**	0%	10%
G	10%	15%	**70%**	5%
S	10%	0%	10%	**80%**

Table 4: *Confusion matrix in percentage of identified speaker's origins using **duration** (average 67.1%)*

	F	I	G	S
F	**95%**	0%	5%	0%
I	0%	**70%**	25%	5%
G	15%	15%	**70%**	0%
S	30%	30%	40%	**0%**

Table 5: *Confusion matrix in percentage of identified speaker's origins using **F0 slope** (average 54.3%)*

	F	I	G	S
F	**55%**	10%	30%	5%
I	30%	**30%**	30%	10%
G	5%	10%	**80%**	5%
S	30%	10%	10%	**50%**

differences can be partly explained by the fact that Spanish speakers are less numerous and less homogeneous because they come from different areas of Spain and so have different Spanish accents. Energy (Table 3) is interesting because it discriminates the four different origins more homogeneously. 65% to 80% of speakers are correctly identified for each origin.

Duration is a good cue for identifying speakers of French origin (cf. Table 4). This is perhaps due to the fact that the lexical accent in French is mainly marked by last syllable lengthening. Figure 1 shows this fact: for the word "seven", the French speakers are the only ones who have the last syllable longer than the first one. We also see in this figure that Spanish and Greek have very similar syllable durations and can not be distinguished.

It is important to notice that the extraction of prosodic parameters is a very difficult undertaking. For instance, F0 estimation, and identifying the duration of the beginning and end of some phones is tricky. Improvement in the extraction of these prosodic parameters would most likely increase the identification rate of the speaker's L1.

5.2. Combination of prosodic parameters and MFCC

In the previous section, we studied independently each prosodic parameter. Now, we wish to evaluate if combining them can allow better identification of foreign accents. Our combined parameters are duration, energy and MFCC. We decided to ex-

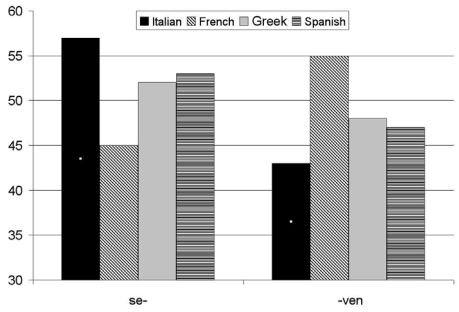

Figure 1: *Duration distribution (in %) for word "seven" in function of speaker's L1 language. HIWIRE corpus.*

Table 6: *Percentage of well identified speaker's origins*

Combined parameters	Ident. rate (%)
MFCC + energy (i)	82.9
MFCC + duration (i)	85.7
MFCC + duration + energy (ii)	87.1

clude the F0 parameter because it achieved poor identification results in the previous experiments. Two approaches have instead been proposed:

- The first approach (i) consists in using two streams: MFCC are used in a first stream and in the other we put duration or energy. Emitting probabilities $b_j(o_t)$ of observation vectors o_t for each frame t are computed according to the following equation:

$$b_j(o_t) = \prod_{s=1}^{S} \left(\sum_{m=1}^{M_{js}} c_{jsm} N(o_{st}; \mu_{jsm}; \Sigma_{jsm}) \right)^{\gamma_s}$$

  M_{js} is the number of Gaussians for the state j and the stream s; c_{jsm} is the weight of the m^{th} Gaussian; $N(o; \mu; \Sigma)$ represents the pdf of observation vector o of mean μ and covariance Σ; γ_s is the weight affected to the stream, default 1.

- For the second approach (ii), two steps are necessary. Firstly, for each parameter and a given speaker, we compute the number of syllables identified for each origin (the same process as in section 4.5). Then we sum up all the results for this given speaker, and the origin which gathers the largest number of syllables is affected to the speaker.

The combination of MFCC and energy (i) gives the same result as MFCC alone (cf. Table 6). On the other hand, the combination of MFCC and duration (i) gives an improvement compared to MFCC alone: 85.7% versus 82.9% but not significant.

The best result is obtained using method (ii) and combining MFCC + duration + energy. Method (ii) gives equal importance to all three parameters. It might be interesting to try different weights for each parameter.

6. Conclusion

In this article, we have proposed an automatic approach for identifying a non-native speaker's origin. This approach is based on acoustic prosodic parameters (duration, energy and the fundamental frequency of syllables). The proposed approach needs the transcription of the utterances. Experiments were conducted on an English speech database pronounced by French, Italian, Greek and Spanish speakers. This approach requires transcription of pronounced utterances.

We have found that duration and energy are promising parameters for correct identification. Moreover, combining these parameters with MFCC brings an error reduction rate of 24% compared to the use of MFCC alone. For duration, it may be interesting to replace the syllable duration by the vocalic nucleus duration in order to avoid duration problems with consonants.

In the future, it may be appropriate to complete the experimental results obtained in this work by an analysis of acoustic realizations (spectrograms, F0 curves, etc.) with the help of a phonetician. It will be interesting to evaluate the impact of replacing manual transcription by automatic recognition. Moreover, the test must be conducted on the larger corpus to confirm the obtained results.

7. Acknowledgments

This work was partially founded by the *HIWIRE* European project (*sixth framework program, IST*).

8. References

[1] L. Arslan and J. Hansen, "Language Accent Classification in American English," in *Speech Communications*, vol. 18, no. 4, 1996, pp. 353–367.

[2] K. Kumpf and R. W.King, "Automatic Accent Classification of Foreign Accented Australian English Speech," in *ICSLP*, 1996, pp. 1740–1743.

[3] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Discriminative Phoneme Sequences Extraction for Non-Native Speaker's Origin Classification," in *ISSPA*, 2007.

[4] C. Horgues, "Contribution à l'étude de l'accent français en anglais. Quelques caractéristiques prosodiques de l'anglais parlé par des apprenants francophones et leur évaluation perceptive par des juges natifs," in *Actes des 8e RJC ED268 Langage et langues*, 2005, pp. 79–83.

[5] F. Ramus and J. Mehler, "Language Identification With Suprasegmental Cues: A Study Based on Speech Resynthesis," vol. 105, no. 1, 1999, pp. 512–521.

[6] J. Hansen and L. Arslan, "Foreign Accent Classification Using Source Generator Based Prosodic Features," in *IEEE ICASSP*, 1995, pp. 836–839.

[7] K. Kumpf and R. King, "Foreign Speaker Accent Classification using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Perception Benchmarks," in *Eurospeech 97*, 1997, pp. 2323–2326.

[8] J.-L. Rouas, M. Barkat-Defradas, F. Pellegrino, and R. Hamdi-Sultan, "Identification automatique des parlers arabes par la prosodie," in *16e JEP*, 2006, pp. 192–196.

[9] D. Fohr and Y. Laprie, "Snorri: an Interactive Tool for Speech Analysis," in *Eurospeech*, 1989, pp. 1613–1616.

[10] T. Kamm, G. Andreou, and J. Cohen, "Experiments in Vocal Tract Length Normalization," in *CAIP workshop : Frontiers in Speech Recognition II*, 1994.