IJMMS 29:2 (2002) 121–124
PII. S0161171202010141
http://ijmms.hindawi.com

INTRODUCTION OF THE SPEAKING RATE IN THE MODEL OF SPEECH RECOGNITION

ABDELLAH YOUSFI and ABDELOUAFI MEZIANE

Received 16 February 2000

We propose an improvement to the centisecond TLHMM model (Meziane, 1999) applied to the sound duration. Indeed, the distribution of the sound duration depends on the speaking rate. An adaptation in a post-processing step is needed. This adaptation is studied by proposing a model of the speaking rate based on average syllabic duration. The experiments elaborated on a set of BDSONS show the interest of this approach. This work is a continuation of those by Meziane (1999) and Suaudeau (1994).

2000 Mathematics Subject Classification: 68T10.

1. Introduction. General phonetic studies consider that the speaking rate of an utterance is reflected in syllable duration [2]. These studies have shown the existence of links between the duration of the phonemes and the global duration of the utterance from which they are extracted [1].

Suaudeau [3] has chosen the average syllabic duration to calculate the speaking rate, and a simple linear model to describe the influence of the speaking rate over the duration of phonemes. This model is used in a post-processing step.

Definition 1.1. Let w be an utterance formed by the phonemes $\phi_{k_1},\ldots,\phi_{k_{\epsilon'}}$, s the number of syllables in this utterance, and $g(\tau)$ the observed duration of the phoneme ϕ_{k_τ}, then the average syllabic duration of this utterance is given by

$$\mathrm{Syl} = \frac{\sum_{\tau=1}^{\epsilon'} g(\tau)}{s}. \tag{1.1}$$

For each phoneme ϕ_{k_τ}, the relation between the duration $g(\tau)$ and the quantity Syl is described by the following linear speaking rate-duration model proposed by Suaudeau [3]:

$$g(\tau) = \alpha_{k_\tau}\mathrm{Syl} + \beta_{k_\tau} + v(\tau) \quad \tau = 1,\ldots,\epsilon, \tag{1.2}$$

where $\alpha_{k_\tau}, \beta_{k_\tau}$ are the parameters of the linear model (these coefficients are estimated on the training set; this estimation gives $\bar{\alpha}_{k_\tau}, \bar{\beta}_{k_\tau}$).

$v(\tau)$ is a noise process with variance $\sigma_{k_\tau}^2$, where $\sigma_{k_\tau}^2$ corresponds to the duration standard variance of the phoneme ϕ_{k_τ}.

2. Use of model (1.2) in a post-processing step. To introduce model (1.2) in the centisecond TLHMM model in a post-processing phase, we have used a procedure based on the following steps.

(1) For an utterance w, the recognition with the centisecond TLHMM model gives a solution w' constituted by phonemes $\phi_{k_1},\dots,\phi_{k_{\epsilon'}}$. If s is the number of syllables in w', there exists an integer i such that the number of syllables composing w is $s+i$, with $-N \le i \le N$ (N is an integer fixed in the experiments). The average syllabic duration of w is, a priori, given by

$$S^i = \frac{\sum_{\tau=1}^{\epsilon'} g(\tau)}{s+i}. \tag{2.1}$$

(2) For each value of the speaking rate S^i, we adjust the parameters of the law of duration in conformity with the model

$$g(\tau) = \bar{\alpha}_{k_\tau} S^i + \bar{\beta}_{k_\tau} + \upsilon(\tau). \tag{2.2}$$

The new parameters of the law of duration are

$$\bar{\mu}^i_{k_\tau} = \bar{\alpha}_{k_\tau} S^i + \bar{\beta}_{k_\tau}, \qquad \bar{\sigma}^i_{k_\tau} = \sigma_{k_\tau}. \tag{2.3}$$

(3) After adjusting the parameters of the centisecond TLHMM model, a recognition by these new parameters is undertaken. For an utterance w, we get $2N+1$ solutions denoted w'_i ($i = -N,\dots,N$), where w'_i is associated with the average syllabic duration S^i.

(4) For these $2N+1$ solutions two cases exist:

(a) The $2N+1$ solutions are identical, we do not do anything.

(b) The $2N+1$ solutions are different. To select one of them, we use the three classic scores [3] (acoustic score, path score, duration score).

As these scores do not give satisfying results, we propose a new score called revised score, and noted $\text{score}_{\text{rev}}$; it combines the three classic scores,

$$\text{score}_{\text{rev}} = \begin{cases} \text{score}_{\text{duration}} & \text{if one of the solutions } w'_i \text{ verifies} \\ & \qquad \text{Delta-dura} < \text{Delta-acou} < \text{Delta-path}, \\ \text{score}_{\text{path}} & \text{if one of the solutions } w'_i \text{ verifies} \\ & \qquad \text{Delta-dura} > \text{Delta-acou} > \text{Delta-path}, \\ \text{score}_{\text{acous}} & \text{else}, \end{cases} \tag{2.4}$$

where

$$\begin{aligned} \text{Delta-acous} &= \frac{\text{score}_{\text{acous}}(w'_i) - \text{score}_{\text{acous}}(w')}{\text{score}_{\text{acous}}(w')}, \\ \text{Delta-dura} &= \frac{\text{score}_{\text{duration}}(w'_i) - \text{score}_{\text{duration}}(w')}{\text{score}_{\text{duration}}(w')}, \\ \text{Delta-path} &= \frac{\text{score}_{\text{path}}(w'_i) - \text{score}_{\text{path}}(w')}{\text{score}_{\text{path}}(w')}, \end{aligned} \tag{2.5}$$

$\text{score}_{\text{acous}}(w')$: acoustic score of the path associated with w',
$\text{score}_{\text{duration}}(w')$: duration score of the path associated with w',
$\text{score}_{\text{path}}(w')$: path score of the path associated with w'.

TABLE 3.1. Rate of errors for the different scores.

	Group 1	Group 2	Group 3
Without duration	4.62%	8.61%	10.56%
Centisecond TLHMM	1.54%	3.06%	4.17%
Acoustic score	1.28%	2.5%	3.61%
Path score	0.85%	2.78%	3.89%
Duration score	2.14%	2.78%	4.17%
Score-rev	2.14%	2.22%	3.89%

2.1. Likelihood of the observation sequence. Let $y_1, y_2, \dots, y_T$ be the observations generated by the centisecond TLHMM model associated with the phonetic sequence

$$(\wedge_i)_{1\le i\le\epsilon} = \{(\phi_{k_i}, \theta_i) \mid i = 1, \dots, \epsilon\}. \tag{2.6}$$

The likelihood of $(y_1, \dots, y_T)$, taking into account the speaking rate S^p, is given by the formula

$$\begin{aligned} &\Pr(y_1, \dots, y_T,\ d_1, \dots, d_\epsilon) \\ &\quad = \sum_{\xi_T} \pi_{i_1} b_{i_1}(y_1) \times \left(\prod_{n=2}^{\theta_2 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \right) \times \varphi_{k_1}^{S^p}(d_1) \\ &\quad\quad \times \left(\prod_{n=\theta_2}^{\theta_3 - 1} a_{i_{n-1} i_n} b_{i_n}(y_n) \right) \times \varphi_{k_2}^{S^p}(d_2) \times \cdots \times \left(\prod_{n=\theta_\epsilon}^{T} a_{i_{n-1} i_n} b_{i_n}(y_n) \right) \times \varphi_{k_\epsilon}^{S^p}(d_\epsilon), \end{aligned} \tag{2.7}$$

where

(i) ξ_T is the path of length T associated with the phonetic sequence $(\wedge_i)_{1\le i\le\epsilon}$.

(ii) $\varphi_k^{S^p}(\cdot)$ is the law of duration of the phoneme ϕ_k taking into account the speaking rate S^p. The mean and the variance of this law are $\bar{\mu}_k^p$ and $\bar{\sigma}_k^p$, respectively.

(iii) θ_τ is the temporal index of the first state of the phoneme ϕ_{k_τ}.

(iv) d_τ is the number of the observations emitted in the phoneme ϕ_{k_τ}, $d_\tau = \theta_{\tau+1} - \theta_\tau$.

3. Experiments. The vocabulary is composed of 20 numbers (0–19) extracted from the database “BDSONS.” The experiments are achieved on three groups of speakers. The first group is composed of 13 male speakers, and the second one of 20 speakers (male, female), the third group is composed of the same speakers as the second group, but with different utterances.

The acoustic parameters are composed of the first 8 Mel frequency cepstral coefficients (MFCC).

For $N = 1$, the results obtained for the different scores are as in Table 3.1.

4. Conclusion. We note that the rate of error in the third group improves when we take into account the speaking rate. We hope to get good results by developing a new score which combines more significantly the different scores. Our model is validated on a corpus formed by utterances not containing enough syllables and pronounced in a normal rhythm. We think that an application on a vocabulary composed of polysyllabic utterances will show significantly the interest of the introduction of this factor in the model of automatic speech recognition.

References

[1] Y. Gong and W. C. Treurniet, *Duration of phones as function of utterance length and its use in automatic speech recognition*, European Conference on Speech Communication and Technology (Berlin), September 1993.

[2] M. Rossi, *l'Intonation: de l'Acoustique à la Sémantique*, GALF Groupe de la Communication Parlée, 1981, pp. 40–53 (French).

[3] N. Suaudeau, *Un modèle probabiliste pour intégrer la dimension temporelle dans un système de reconnaissance automatique de parole* [*A probabilistic model to integrate temporal dimension in an automatic system of recognition of word*], Thèses parole avec résumés, Université Rennes I, March 1994.

Abdellah Yousfi: Département de Mathématique, Faculté des Sciences, Université Mohamed Premier, Oujda, Morocco
E-mail address: yousfi.abdellah@sciences.univ-oujda.ac.ma

Abdelouafi Meziane: Département de Mathématique, Faculté des Sciences, Université Mohamed Premier, Oujda, Morocco
E-mail address: meziane@sciences.univ-oujda.ac.ma

Journal of Applied Mathematics and Decision Sciences

Special Issue on Decision Support for Intermodal Transport

Call for Papers

Intermodal transport refers to the movement of goods in a single loading unit which uses successive various modes of transport (road, rail, water) without handling the goods during mode transfers. Intermodal transport has become an important policy issue, mainly because it is considered to be one of the means to lower the congestion caused by single-mode road transport and to be more environmentally friendly than the single-mode road transport. Both considerations have been followed by an increase in attention toward intermodal freight transportation research.

Various intermodal freight transport decision problems are in demand of mathematical models of supporting them. As the intermodal transport system is more complex than a single-mode system, this fact offers interesting and challenging opportunities to modelers in applied mathematics. This special issue aims to fill in some gaps in the research agenda of decision-making in intermodal transport.

The mathematical models may be of the optimization type or of the evaluation type to gain an insight in intermodal operations. The mathematical models aim to support decisions on the strategic, tactical, and operational levels. The decision-makers belong to the various players in the intermodal transport world, namely, drayage operators, terminal operators, network operators, or intermodal operators.

Topics of relevance to this type of decision-making both in time horizon as in terms of operators are:

- Intermodal terminal design
- Infrastructure network configuration
- Location of terminals
- Cooperation between drayage companies
- Allocation of shippers/receivers to a terminal
- Pricing strategies
- Capacity levels of equipment and labour
- Operational routines and lay-out structure
- Redistribution of load units, railcars, barges, and so forth
- Scheduling of trips or jobs
- Allocation of capacity to jobs
- Loading orders
- Selection of routing and service

Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://www.hindawi.com/journals/jamds/guidelines.html. Prospective authors should submit an electronic copy of their complete manuscript through the journal Manuscript Tracking System at http://mts.hindawi.com/, according to the following timetable:

Manuscript Due	June 1, 2009
First Round of Reviews	September 1, 2009
Publication Date	December 1, 2009

Lead Guest Editor

Gerrit K. Janssens, Transportation Research Institute (IMOB), Hasselt University, Agoralaan, Building D, 3590 Diepenbeek (Hasselt), Belgium; Gerrit.Janssens@uhasselt.be

Guest Editor

Cathy Macharis, Department of Mathematics, Operational Research, Statistics and Information for Systems (MOSI), Transport and Logistics Research Group, Management School, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium; Cathy.Macharis@vub.ac.be