AN EXPERIMENT IN VOCAL TRACT LENGTH ESTIMATION

A.Sitchi[1], F. Grenez[1], J. Schoentgen[1,2]
[1]Laboratory of Images, Signals and Telecommunication Devices, Université Libre de Bruxelles, Bruxelles, Belgium
[2]National Fund for Scientific Research, Belgium

Abstract: **The presentation concerns the estimation of the vocal tract length of a speaker on the base of her formant frequencies and the formant frequencies and known tract length of a reference speaker. The length prediction is founded on a rule inferred from Webster's equation that describes the propagation of a planar acoustic wave in a loss-less vocal tract. The length prediction experiments have been cross-language, cross-gender and cross-corpora. Results show that the relative length prediction error is less than 3%, which is inferior to the error made assuming typical tract lengths of 15 and 17 cm for male and female speakers respectively.**

I. INTRODUCTION

This study is devoted to the estimation of the vocal tract length of a speaker by means of his formant frequencies and the formant frequencies and default tract length of a reference speaker.

Several studies have been devoted to the topic of tract length estimation, because one may argue that the tract length is an anatomical cause of inter-speaker variability [2]. Possible applications of predicting tract lengths from acoustic data are speaker normalization and the facilitation of acoustic-to-articulatory inversion [1]. A majority of studies have focused on length normalization with a view to achieving speaker normalization, without attempting to estimate the tract length explicitly.

The default length is the vocal tract length with lips and larynx in neutral positions. Lip rounding or spreading and larynx raising or lowering are phonetically relevant gestures that mark vowel timbre and which overlay a speaker's anatomically conditioned default length.

Several methods have been used to estimate the vocal tract length from speech. One is based on a known formula that relates the length of a uniform loss-less acoustic tube to its natural frequencies when the tube is open at one end and closed at the other. The vocal tract length is estimated by averaging several length values obtained by means of several observed formants [3], [4].

Paige et *al.* have proposed to estimate the tract length using low-order poles and zeros of the lip impedance, omitting the assumption of uniform cross-sections [5]. The lip impedance poles correspond to the natural frequencies of the tract closed at both ends, which cannot be measured from the speech signal directly. Be that as it may, Paige's approach has in common with [1] that it aims at obtaining length estimates on the base of acoustic data only.

The method we have investigated enables estimating the unknown vocal tract length of a speaker by means of his formant frequencies as well as the known vocal tract length and formant frequencies of a reference speaker. The experiments that have been carried out include predicting tract lengths across genders and linguistic communities. The focus has been on default tract lengths, because the general framework has been acoustic-articulatory inversion, which assumes the default lengths to be known and the deviations therefrom to be computable.

II. METHODS

The method is based on an observation, made by Ungeheuer, concerning Webster's equation, which describes the propagation of planar loss-less acoustic waves in non-uniform ducts [7]. Webster's equation suggests that when the longitudinal dimension of an acoustic tube is multiplied by a constant, its natural frequencies change inversely proportional to that same constant. Applying this observation to the vocal tract would suggest that multiplying the length of the vocal tract by a number causes the formants frequencies to be divided by the same number. Mol, for instance, has tested this prediction by means of the Peterson and Barney data [8] by displaying the first and second formant averages for men, women and children in a chart and observing that the averages are positioned for each vowel on a straight line through the chart origin [6].

A. Estimation of the factor of proportionality

Because the first three formants of all vowels are assumed to obey the rule of inverse proportionality, one calculates as follows the multiplicative constant α, which is assumed to explain inter-speaker formant differences owing to default length differences.

Models and analysis of vocal emissions for biomedical applications: 5th international workshop: December 13-15, 2007: Firenze, Italy, ed. by C. Manfredi,
ISBN 978 88-8453-673-3 (print) ISBN 978-88-8453-674-7 (online)

$$\alpha = \frac{\sum_{i=1}^{N} F_{i,ref}}{\sum_{i=1}^{N} F_i} \qquad (1)$$

Symbol $F_{i,ref}$ designates the formants frequencies of a set of vowels of a reference speaker whose average vocal tract length is known and F_i the formant frequencies of a set of vowels of the speaker whose vocal tract length is unknown. Symbol N equals the number of formants per vowel time the number of vowel categories. It is desirable that the vowel categories and the number *N* of formant frequencies are identical for the target and reference speakers, because of the vowel-typical vocal tract lengthening and shortening that must be averaged out when the goal is the estimation of the default length.

Once factor of proportionality, α, has been obtained, the unknown tract length *L* can be estimated via the known tract length of the reference speaker.

$$L = \alpha L_{ref} \qquad (2)$$

B. Corpora

Corpora are divided into reference and test corpora. The first and the second reference corpora comprise the vocal tract lengths and first three formant frequencies of 10 French vowels sustained each by 4 speakers (2 males and 2 females) [9] and one male speaker [10] respectively). Hereafter, these speakers are labeled MS_1, MS_2, FS_1, FS_2 and MS_3.

A third reference corpus comprises the tract lengths and first three formant frequencies for 10 American-English vowels sustained by one female speaker [11] (labeled FS_{AE}).

The formant frequency data published in the framework of these corpora have been obtained via measured vocal tract cross-sections and lengths combined with acoustic models. The purpose has been to guarantee the best possible match between published acoustic and morphological data.

This is, however, problematic when the objective is to test relations (1) and (2) because for these corpora the formant frequency data cannot be assumed to be independent of the model the predictions of which they are expected to validate.

Therefore, only those corpora have been retained as test corpora for which the formant frequencies have been determined from the speech spectra directly, loose from any Webster's equation-based modeling. One test corpus comprises the tract lengths and formant frequencies measured for one male speaker who has sustained 10 American-English vowels [11]. The second test corpus comprises the vocal tract lengths and three formant frequencies of 5 Russian vowels produced by one male speaker [12]. The American English and Russian speakers are hereafter labeled MS_{AE} and MS_R respectively.

The area functions and lengths published in [9] and [11] have been obtained by nuclear resonance imaging. The cross-sections and lengths in [12] are the well-known Russian vowel data published by G. Fant. They have been compiled on the base of X-ray images. The shapes and lengths published in [10] have been recorded by a combination of phonetic a priori knowledge, visual inspection of human speakers and X-ray imaging.

The default length for each speaker has been obtained by averaging the vowel-typical lengths.

III. Results

A. Experiment 1

The experiment consists in predicting the vocal tract length of American-English test speaker MS_{AE} by means of each reference speaker in turn. Table 1 shows the length prediction results. One sees that the absolute maximum relative error is less than 2 %.

Table 1: Relative error in % and proportionality factors α obtained for American-English male test speaker MS_{AE}. Symbol *L* is the measured default length.

	Test: MS_{AE} L = 17.14cm	
References	α	Relative error (%)
MS_1	0,93	-0,15
MS_2	0,93	-0,54
FS_1	1,08	-1,77
FS_2	1,03	0,91
MS_3	0,96	-0,91
FS_{AE}	1,25	-0,06

B. Experiment 2

The experiment consists in predicting the vocal tract length of Russian test speaker MS_R by means of each reference speaker in turn. This experiment has involved five of Fant's Russian vowels [12]. The number of

vowels has been the same for all speakers. The Russian and reference vowel qualities have been chosen to be as similar as possible.Table 2 shows the length prediction results. One sees that the absolute maximum relative error is less than 2.5 %.

Table 2: Relative error in % and proportionality factors α obtained for Russian male test speaker MS_R. Symbol L is the measured default length.

	Test: MS_R L = 17.6cm	
References	α	Relative error (%)
MS_1	0.95	-1.12
MS_2	0.96	0.13
FS_1	1.13	-2.41
FS_2	1.08	-0.29
MS_3	1.01	-0.96
FS_{AE}	1.27	-0.47

C. Experiment 3

The experiment involves speakers MS_R and MS_{AE} as test and reference speakers respectively. Then the proportionality factor α is equal to 1.04 and the relative error equal to -1.18 %. Inverting the roles of speakers MS_R and MS_{AE} gives rise to the same relative error in absolute value because relation (2) shows that estimating one length from another and vice versa means replacing constant α by 1/α and the relative error by its negative.

D. Experiment 4

This experiment has been carried out with the six speakers originally assigned to the reference corpora. Within this experiment, each speaker has been given in turn the role of "reference" speaker from whom the lengths of the other five speakers are predicted. Table 3 reports the proportionality factors α above the main diagonal and the relative error in percent below the main diagonal. The line indexes refer to "reference" and the column indexes to "test" speakers. In Table 3, the maximum relative error is less than 3% whoever the "reference" speaker.

One should keep in mind that predicting the lengths of speakers belonging to these corpora is a necessary, but not sufficient, test. The reason is that for these speakers the formant frequencies have not been obtained independently of Webster's equation relation (2) is a consequence of.

Table 3: Relative error in % (below the diagonal) and proportionality factors α (above the diagonal) for 6 speakers [9,10,11], each taking the role of "reference" speaker in turn. Symbol L is the measured default length in cm.

	MS_1 L=18,54	MS_2 L=18,24	MS_3 L=18	FS_1 L=16,01	FS_2 L=16,42	FS_{AE} L=13,73
MS_1		0,99	0,96	0,85	0,9	0,74
MS_2	0,69		0,97	0,86	0,9	0,75
MS_3	-0,76	-1,46		0,88	0,93	0,77
FS_1	-1,62	-2,33	-0,86		1,05	0,87
FS_2	1,06	-0,38	1,81	2,64		0,83
FS_{AE}	0,09	-0,6	0,85	1,68	-0,98	

E. Comparison with standard assumptions

Often one assumes that the standard vocal tract length for men is 17 cm and for women 15 cm. One question is whether predicting the tract lengths by means of formant frequencies and the data of a reference speaker causes relative errors that are smaller than those that would have been obtained by making the above default assumptions. One sees in Table 4 that these assumptions cause relative errors between -9.2% and +8.6%. The (absolute) average is 6.1%, which must be compared to the average of 0.72% of Table 1 and 0.90% of Table 2. Table 4 therefore suggests favouring length prediction over length standardization via default values.

Table 4: Relative length error in % using standard tract lengths of 17 and 15 cm for males and females respectively. Symbols L_{OBS} and L_{STD} designate the observed length and the standard length respectively in cm. Symbol ε designates the relative error in %.

	MS_1	MS_2	FS_1	FS_2	MS_3	MS_{AE}	FS_{AE}	MS_R
L_{OBS}	18,54	18,24	18	16,01	16,42	17,14	13,73	17,6
L_{STD}	17	17	17	15	15	17	15	17
ε	8,31	6,8	5,56	6,31	8,65	0,82	-9,25	3,41

F. Correlation between factors of proportionality α and length prediction errors

Relation (2) is applicable to arbitrary test and reference lengths, whatever the length difference. Relation (2) therefore predicts that no correlation is expected between calculated length errors and constants of proportionality α. For the grouped Experiments 1 and 2 and for

Experiment 4, the correlations between calculated lengths errors and factors of proportionality are -0.2494 and -0.1493 respectively. These are not statistically significant.

IV. DISCUSSION AND CONCLUSION

a) Results suggest that estimating unknown tract lengths via measured formant frequencies and a reference tract length is a valid method that causes errors that are smaller than those made by assigning standard lengths to male and female tracts.
b) The different experiments have involved cross-linguistic & cross-gender length predictions. The results suggest that these cross-factor predictions do not cause length estimation errors to be larger than within-factor predictions. A possible explanation is that observed errors are the combined effect of measurement errors (morphological and acoustic), the disparity between the recording conditions of acoustic and length data (which may not have been simultaneous) as well as the disagreement between predicted and recorded data, and that these combined errors are larger than the average errors caused by cross-linguistic vowel category or gender mismatch.
c) Length estimation errors and factors of proportionality α are not statistically significantly correlated. This is an indirect test of the validity of relation (2). Indeed, if relation (2) were a crude approximation only of an unknown relation between the vocal tract lengths of two speakers, one would expect to observe increasing length estimation errors with increasing factors of proportionality. The reason is that linear relation (2) is then expected to approximate that link the better the smaller the difference between the reference and test tract lengths. The lack of observed correlations suggests, however, that identity (2) is a valid approximation of the relation between the default vocal tract lengths of two speakers, whatever the difference in vocal tract size, as long as up to three formants are involved in the comparison.

ACKNOWLEDGEMENTS

We acknowledge support of FET project “Audio-visual to articulatory speech inversion” (ASPI) and of COST Action 2103 “Advanced voice assessment”.

REFERENCES

[1] S. Dusan, L. Deng, “ Vocal tract length normalization for acoustic-to-articulatory mapping using neural networks,” *J. Acoust. Soc. Am,* vol. 106, pp. 2181, 1999.

[2] D. Paczolay, A. Kocsor and L. Toth, “ Real-time vocal tract length normalization in a phonological awareness teaching system,” *Lect. notes comput. Sc Springer Verlag,* vol. 2807, pp. 309-314, 2003.

[3] S. Dusan, “Estimation of speaker’s height and vocal tract length from speech signal,” *Proc. Inter-. Speech* Lisbon, pp. 1989-1992, 2005.

[4] B.F. Necioglu, M.A. Clements and T.P. Barnwell, “Unsupervised estimation of the human vocal tract length over sentence level utterances,” *Proc. IEEE ICASSP* Istanbul, pp. 1319-1322, 2000.

[5] A. Paige, V.W. Zue, “Calculation of vocal tract length,” *IEEE Transactions on audio and electroacoustics,* vol. 18, pp. 268-270, 1970.

[6] H. Mol, *Fundamentals of Phonetics*, The Hague, Nederlands: Mouton, 1970.

[7] G. Ungeheuer, *Elemente einer akustischen Theorie der Vokalartikulation*, Germany: Springer Verlag, 1962.

[8] G. E. Peterson, H. L. Barney, “Control methods used in a study of the vowels,” *J. Acoust. Soc. Am*, vol. 24, pp. 175-184, 1952.

[9] M. George, *Analyse du signal de parole par modélisation de la cinématique de la fonction d’aire du conduit vocal*, Bruxelles : Université Libre de Bruxelles, 2001, pp. 177-178.

[10] M. Mryayti, *Contributions aux études sur la parole*. France : Institut National Polytechnique de Grenoble, 1976.

[11] B. Story, I. Titze and E. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” *J. Acoust. Soc. Am,* vol. 100, pp. 537-554, 1996.

[12] G. Fant, *Acoustic theory of speech production*, The Hague, Nederlands: Mouton, 1970, pp.109.