# Vocal Tract Normalization Based on Formant Positions

**Nikša Jakovljević*, Dragiša Mišković*, Milan Sečujski*, Darko Pekar†**

* Faculty of Engineering, Trg Dositeja Obradovića 6, Novi Sad, Serbia
{jakovnik, dragisa, secujski}@uns.ns.ac.yu
†Alfanum Ltd., Trg Dositeja Obradovića 6, Novi Sad, Serbia
darko.pekar@alfanum.co.yu

**Abstract**

This paper presents our initial results in a new approach to vocal tract normalization (VTN). In experiments based on continuous automatic speech recognition (ASR) the VTN procedure is in general carried out in both training and test phase. In the training phase it is used to obtain speaker independent acoustic models of phones. In the test phase it is used to convert input observations into observations nearer to the ones corresponding to the universal speaker. The approach described in this paper is new, because instead of training a single set of acoustic models for the universal speaker, several sets of acoustic phone models corresponding to speakers with similar vocal tract lengths were created. Instead of using the VTN procedure in the test phase, the recognized sequence estimated as the most likely one among sequences based on different acoustic model sets was identified as the final recognition result.

**Normiranje vokalnega trakta na podlagi lege formantov**

V prispevku so predstavljeni začetni rezultati novega pristopa k normiranju vokalnega trakta (NVT). V eksperimentih, ki temeljijo na samodejnem razpoznavanju tekočega govora, se postopek NVT izvaja tako v učni kot v testni fazi. V učni fazi se uporablja za pridobivanje akustičnih modelov fonov, ki niso odvisni od govorca. V testni fazi se uporablja za pretvorbo vhodnih opažanj v opažanja, ki so bližja tistim, ki ustrezajo univerzalnemu govorcu. Pristop, ki je opisan v tem prispevku, je nov: namesto da bi učili posamezno množico akustičnih modelov za univerzalnega govorca, je bilo ustvarjenih več množic akustičnih modelov fonov, ki ustrezajo govorcem s podobno dolžino vokalnega trakta. Namesto uporabe postopka NVT v testni fazi je bil končni rezultat razpoznavanja prepoznano zaporedje, ki je bilo ocenjeno kot najbolj verjetno med zaporedji, temelječimi na različnih množicah akustičnih modelov.

## 1. Introduction

Most of today's automatic speech recognition (ASR) systems are based on hidden Markov models (HMM). Acoustic variations between training and test conditions, caused by different microphones, channels, background noise as well as speakers, are known to deteriorate ASR performance. Speaker variations can be divided into extrinsic and intrinsic. Extrinsic variations are related to cultural variations among speakers as well as their emotional state, resulting in diverse speech prosody features. Intrinsic variations are related to speaker anatomy (vocal tract dimensions) and they manifest in different formant positions of a given phoneme. Procedures for reducing variation caused by different vocal tract dimensions in feature domain are known as vocal tract normalization (VTN) procedures, whereas procedures in acoustic model domain are referred to as adaptation procedures.

In this paper the improvements of the AlfaNum ASR system obtained by the VTN procedure will be presented. In section 3, a description of the corpus and features used is given. Section 4 contains a description of HMM modeling on phonetic level. A description of VTN procedures is given in section 5. Experiment results are presented in section 6, followed by conclusions in section 7.

## 2. Goal of the paper

Variations in vocal tract length are the main reason for diverse formant positions within a given phoneme spoken by different persons, hence formant based spectrum warping is more than reasonable. Unfortunately, this approach to VTN has several disadvantages: (i) formant positions are context dependent and could vary largely with different context even for a single speaker; (ii) there are overlaps between different formants across vowels spoken by various speakers; (iii) existing formant estimation techniques are not robust enough. Zhan and Waibel (1997) showed that VTN based on formant positions did not result in any performance improvement, since formant frequency could not reflect difference in vocal tract length among speakers because they are calculated with an unconstrained context and there is no guarantee of phone balance in context among speakers (Zhan, Waibel, 1997). Exact phone boundaries can be used to avoid the problem of context dependency of formant positions only in the training phase, since they are not known in the test phase. In this approach exact phone boundaries are used in training phase to make clusters of speakers with similar vocal tract lengths. For each cluster of speakers a set of acoustic models is created. In the test phase the recognized sequence estimated as the most likely one among sequences based on different acoustic model sets was identified as the final recognition result. Division of the training set into subsets i.e. speaker clusters would reduce the number of utterances per cluster and decrease robustness of consequent acoustic models. In order to overcome this problem a warping procedure was used to extend each training subset with utterances spoken by speakers out of the cluster.

## 3. Database and features

The used corpus is a part of the Serbian SpeechDat database (Đurić, Pekar, Jovanov, 2002), containing only utterances spoken by male speakers. The corpus in this experiment is reduced only to those speakers for which at least 10 instances of each vowel could be found in the database in order to achieve good vocal tract length estimation for each speaker in the corpus. The Serbian SpeechDat database was recorded through the public switched telephone network and sampled at 8 kHz with 8-bit A-law quantization. The training set contains 14496 utterances spoken by 340 speakers. For testing system

performance 2 test sets were used. The first test set contains 184 utterances spoken by 17 different speakers. No utterance spoken by any of these speakers is present in the training set. The second test set contains 435 utterances spoken by 17 different speakers. Some of the utterances spoken by these speakers are present in training set but not the same ones. The feature vector which was used consists of 2 streams. The first stream contains 6 energy coefficients: normalized energy, logarithm of the energy and their first and second derivatives. The second stream contains 36 coefficients (12 static, 24 dynamic), which describe spectral envelope and its changes in time. These 12 static coefficients describe spectral slopes, or more precisely, differences in energy between successive filter banks. Filter banks divide the Mel-scaled spectrum from 50 to 3800 Hz into 27 regions of equal width. Slopes are evaluated for every other filter bank starting from the third one. Spectral components below 300 Hz and above 3400 Hz are given less relative importance because the AlfaNum ASR system uses telephone quality recordings where these components are distorted. The feature vector is estimated on 30 ms long segment. Overlapping between successive segments is 20 ms.

## 4. Models

For the purposes of this experiment, several changes into the phonetic inventory of the Serbian language had to be introduced. Instead of the standard 5 vowels in Serbian, two sets containing 5 long and 5 short vowels are taken into consideration (the boundary between the two being 65 ms), and the phone /ə/ (IPA notation) is regarded as a standard vowel as well. The distinction based on vowel length is motivated by a need to model steady formant positions within long vowels better. Closure and explosion of affricates and stops are modelled separately and referred to as subphones. The basic modelling unit is a context dependent phone or subphone referred to as *triphone*. Silence and non-speech sounds present in the corpus are modelled as context independent units.

The number of states per model is proportional to the average duration of all the instances of the corresponding phone in the database. The number of mixtures per state depends on the distribution of observations in the feature space and is determined dynamically. During the initial training the maximum number of mixtures and the minimum number of observations per mixture are specified.

Using triphones instead of monophones leads to a very large set of models and insufficient training data for each triphone. All HMM state distributions would be robustly estimated if sufficient observations were available for each state. This could be achieved by extending the training corpus or by including observations related to acoustically similar states. The second solution was chosen as being less expensive, even though it generates some suboptimal models. More details about the tying procedure used can be found in (Jakovljević, Pekar, 2005).

## 5. Formant estimation and the warping function

Variations in vocal tract length are the main reason for diverse formant positions within a given phoneme spoken by different persons, therefore formant based spectrum warping is more than reasonable. Unfortunately, the existing formant estimation techniques are not robust enough. Some of the most frequent errors are: formant merging, shifting formant frequencies towards harmonics and false maximum caused by channel distortion (Gouvea, 1998). The algorithm used for formant detection was the one described in (Welling, Ney, 1998). The algorithm does not perform sufficiently well for Serbian vowels /u/ and /i/. The first and the second formant of the vowel /u/ are in many cases very close to each other in the spectrum, and the algorithm can erroneously identify them as a single formant, thus the third formant is detected as the second. The first formant of the vowel /i/ is very low and in some cases attenuated by the channel, and the algorithm often identifies the peak in the range between 600 and 1800 Hz as the first formant. This kind of error is caused by pre-emphasis, but omitting pre-emphasis would result in wrong formant positions for other vowels. Coarticulation is known to cause formant transition in vowels. If the vowel is too short, positions of its formants cannot reach context neutral values. In order to reduce this type of variability, formant position estimation is based on the most reliable 50% of the frames of long vowels /a/, /e/ and /o/, which are those in the middle of the vowel. The results published show minor differences in performance for various VTN function types (Zhan, Westphal, 1997; Uebel, Woodland, 1999; Pitz, 2005). The linear function was chosen as the simplest one and applied in addition to the Mel-scale warping mentioned above. The most natural way to evaluate the frequency warping factor is as a mean value of the ratio of the universal and the current formant value, the universal formant value being the mean formant value for a given phone across all speakers. The frequency warping factor $\alpha_c$ (i.e. linear function slope) for a given speaker can thus be estimated as follows:

$$\alpha_c = \sum_i \sum_f \frac{\mu_{il}}{F_{ilf}} \tag{1}$$

where $\mu_{il}$ is the mean value of the $i$-th formant in the phone $l$ across all speakers, and $F_{ilf}$ is the current value of the $i$-th formant in the frame $f$ of the phone $l$. This approach to warping factor estimation does not consider the possibility of false formant estimation. A more robust way to estimate $\alpha_c$ is as follows:

$$\alpha_f = \arg\max_{\alpha} \left\{ \prod_i P\left\{ \alpha F_{ilf} \mid il \right\} \right\} \tag{2}$$

$$\alpha_f = \frac{\sum_i F_{ilf}\, \mu_{il} \,/\, {\sigma_{il}}^2}{\sum_i \left( F_{ilf}\, \mu_{il} \,/\, \sigma_{il} \right)^2} \tag{3}$$

$$\alpha_c = \frac{\sum_f \alpha_f \prod_i P\{\alpha_f F_{ilf} \mid il\}}{\sum_f \prod_i P\{\alpha_f F_{ilf} \mid il\}} \tag{4}$$

where $\alpha_f$ is the warping factor for the $f$-th frame, $F_{ilf}$ is the value of the $i$-th formant in the $f$-th frame, $\mu_{il}$ is the mean value of the $i$-th formant in the phone $l$, $s_{il}$ is the standard deviation for the $i$-th formant in the phone $l$, $P\{\alpha_f F_{ilf}|i,l\}$ is the probability that frequency $\alpha_f F_{ilf}$ is actually the $i$-th formant in the phoneme $l$, and $\alpha_c$ is the warping factor for a given speaker. In the first stage for each frame, warping factor $\alpha_f$ is evaluated as most probable warping factor for given vowel (Eq. 2). Under assumption that formant distribution across all speakers for a given vowel is Gaussian, Eq. 2 becomes Eq. 3. In the second stage the warping factor for a given speaker is calculated as the average value across all frames. Taking probability $P\{\alpha_f F_{ilf}|i,l\}$ into account reduces formant

estimation errors. This method could be performed only in the training phase, when phones and theirs boundaries are known. If the warping factor were calculated based on formant positions of only one formant of a single vowel, the reliability factor $P\{\alpha_f F_{ilf}|i,l\}$ would be eliminated, reducing Eq. 4 to Eq. 1.

## 6. Experiments

### 6.1. Finding optimal features for warping factor estimation

The first step of the experiment was finding optimal features for warping factor estimation. The search space contains different combinations of the first 3 formants (F1, F2 and F3) of the vowels /e/, /a/ and /o/. During the evaluation of the formant estimation algorithm, vowels /i/ and /u/ were identified as unreliable (about 40% of observed frames were incorrect). Instead of using an existing ASR system as a reference, a new one using the training corpus adapted for VTN purposes and described in section 2 was trained. Results are thus made independent of the training corpus and none of the utterances of speakers whose utterances are present in the test set are used for acoustic models training. The grammar consists of 195 different words where 8 of them are not present in the VTN test set but are phonetically similar to some of the existing ones. The testing was carried out in a supervised mode to avoid errors caused by incorrect vowel recognition, because the aim of this step was to find optimal features (the set of formants) for reliable warping factor estimation and not to implement VTN procedure itself. In supervised mode phone boundaries are located

| formant | vowel | false | ins | del | WER[%] |
|---|---|---|---|---|---|
| reference system | | 47 | 37 | 0 | 20.90 |
| F2 | /e/ | 39 | 24 | 1 | 15.92 |
| F2 | /a/ /e/ | 38 | 27 | 1 | 16.42 |
| F1 F2 F3 | /a/ | 41 | 31 | 0 | 17.91 |
| F3 | /e/ | 43 | 29 | 1 | 18.16 |
| F2 F3 | /e/ | 42 | 30 | 1 | 18.16 |
| F2 F3 | /a/ /o/ | 44 | 30 | 0 | 18.41 |
| F2 | /a/ /e/ /o/ | 39 | 35 | 1 | 18.66 |
| F2 | /a/ | 46 | 31 | 0 | 19.15 |
| F2 F3 | /a/ /e/ /o/ | 43 | 34 | 1 | 19.40 |
| F3 | /o/ | 46 | 33 | 1 | 19.90 |
| F2 | /e/ /o/ | 44 | 35 | 1 | 19.90 |
| F2 F3 | /a/ | 49 | 32 | 0 | 20.15 |
| F1 | /a/ | 47 | 34 | 1 | 20.40 |
| F2 F3 | /a/ /e/ | 45 | 37 | 1 | 20.65 |
| F2 F3 | /e/ /o/ | 49 | 35 | 0 | 20.90 |
| F2 | /a/ /o/ | 47 | 37 | 0 | 20.90 |
| F3 | /a/ | 46 | 37 | 1 | 20.90 |
| F1 | /e/ | 50 | 43 | 1 | 23.38 |
| F2 F3 | /o/ | 55 | 56 | 0 | 27.61 |
| F2 | /o/ | 64 | 55 | 2 | 30.10 |
| F1 | /o/ | 71 | 52 | 0 | 30.60 |

Table 1: System performances for different features for warping factor estimation

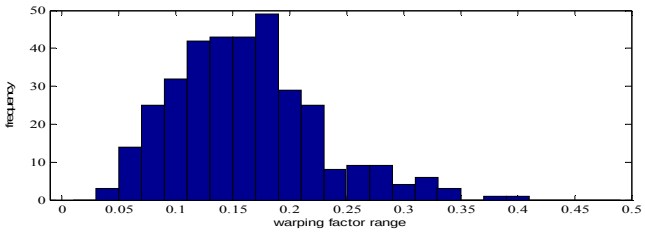


Figure 1. The histogram of the warping factor range

manually. After appropriate warping factor evaluation for each speaker in the test set, the recognition is performed. The results of this phase of the experiment are presented in Table 1. The best system performance is achieved by warping factor estimation based on the second formant of the vowel /e/. Very similar performance is obtained if the warping factor is estimated based on the second formant of vowels /e/ and /a/ instead. Performance improvement comes mostly as a result of a decrease in the number of insertions. Reduction of Eq. 4 to Eq. 1 had no effect since reliability of formant estimation for phoneme /e/ is high. Since the vowel /a/ is the closest one to the neutral vowel /ə/, where each formant frequency is inversely proportional to vocal tract length, it was expected that the VTN based on the formants of the vowel /a/ would produce the best results. However, this was not the case. The results obtained for the formants of the vowel /a/ show some interesting features. If a single formant (F1, F2 or F3) were used for warping factor estimation, or a combination of F2 and F3, the gain is far less than if all 3 formants (F1, F2 and F3) of the same vowel were used. A possible explanation is that during warping factor estimation based only on one formant of a single vowel, warping factor estimation is less reliable, as explained in section 5. It can be seen that experimental results are not very consistent. The system performance in case warping factor estimation is based on F2 of the vowel /a/ is somewhat inferior to the system performance in case the estimation is based on F3 of the vowel /e/. On the other hand, the system with warping factor estimation based on F2 of vowels /e/ and /a/ performs significantly better than the system with estimation based on F2 and F3 of the vowel /e/. One can find further such examples in Table 1. The first formant turned out to be the least appropriate feature for warping factor estimation. A system with warping factor estimation based only on the first formant shows serious degradation of performance in comparison with the referent system in most cases, except for the vowel /a/. In the experiments described in (Gouvea, 1998), the system using warping factor estimation based on the first formant showed the least improvement, but the result was still better than if no VTN procedure had been used. For this reason the first formant was not used in any of the experiments, except in the case of the vowel /a/, because the ratio of its first three formant frequencies is always near to 1:3:5 and it can be shown that F1 contributes to the reliability of estimation of F2 and F3. The second formant has turned out to be the best feature for warping factor estimation. That was not unexpected, since it is known that professional impersonators move their F2 closer to the one of the target speaker, as it seems to be a very important feature of human speaker recognition (Blomberg, Elenius, Zetterholm, 2004). Unfortunately, variations among warping factors obtained in different ways are rather high. Fig. 1 shows the distribution of the

differences between the maximum and minimum warping factor values for each speaker. This is the reason why for some feature combination VTN procedure did not result in any improvement.

### 6.2. Vocal tract normalization

A common method for improving performance is to use separate acoustic model sets for male and female speakers. Instead of creating an universal acoustic model set, 3 separate model sets were created, representing phones uttered by male speakers only. We intend to extend this approach to the models for female speakers. Any division of the training set into subsets would reduce

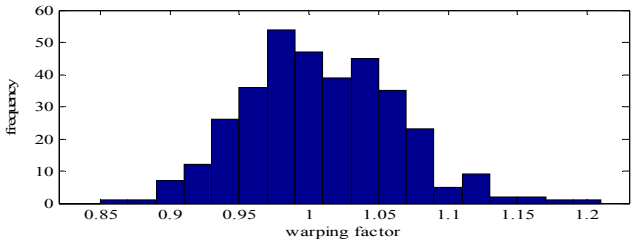


Figure 2. Warping factor histogram for warping factor estimation based on the 2nd formant of /e/

the number of utterances per subset and decrease robustness of consequent acoustic models. In order to overcome this problem warping factors based on F2 of the vowel /e/ were used to extend each of the training sets. The histogram of warping factors for all speakers in the training set is shown in Fig 2. It can be seen that the best coverage of speakers in the training set can be obtained if the subsets of speakers with warping factor values 0.95, 1 and 1.05 are chosen. In order to be able to include the utterances with an inappropriate warping factor in the training set, the spectrum of each such utterance should be scaled with the ratio of its own warping factor and the target warping factor.

The comparative performance is presented in table 2. In this experiment the most successful set of acoustic models describing phones uttered by male speakers was used as the referent system. The referent system was trained on all sentences in the corpus, not only those of speakers for which at least 10 instances of each vowel were found in the database. The test set is the same as the standard set for system evaluation described in section 2. It contains 597 utterances with 735 words spoken by 100 speakers. The grammar consists of 110 words with 40 of them not present in the training set.

In this way the complexity increased 2.8 times (somewhat less than 3 because each subset in the extended VTN system had fewer mixtures than the referent system itself), and the relative improvement is about 7%. It is expected that extension of this approach will result in smaller complexity increase since some of the male and female phone utterances overlap regarding formant positions. The extension of the test set will improve WER resolution, which may give a better picture of the relative improvement.

| system | false | ins | del | WER[%] |
|---|---|---|---|---|
| referent | 39 | 37 | 1 | 8.35 |
| extended VTN | 30 | 26 | 1 | 7.75 |

Table 2: System performance

## 7. Conclusion

In this paper a new approach to the vocal tract normalization procedure is presented. Three separate acoustic model sets are created to describe phones uttered by male speakers only. The utterances are split into 3 classes according to speaker vocal tract length estimated based on F2 of the vowel /e/. Reduction of the number of instances caused by this procedure is overcome by recalculation of warping factors for each utterance. Each model set is trained on the same utterances, but warping factors for an utterance may vary depending on the model being trained. Such an approach omits warping factor calculation during the test procedure, but increases model complexity about 3 times. Achieved relative improvement in WER of 7 % is very small considering the increase in complexity. It is expected that the extension of this approach to acoustic models of phones spoken by female speakers will result in a more significant improvement in performance without such an increase in complexity. On the other hand, this extension will expand the training corpus with utterances spoken by female speakers.

## 8. Acknowledgment

This work was supported in part by the Ministry of Science and Environment Protection of Serbia within the Project “Development of speech technologies in Serbian and their application in ‘Telekom Srbija’” (TR-6144A).

## 9. References

Blomberg. M. D. Elenius, E. Zetterholm, 2004. Speaker verification scores and acoustic analysis of a professional impersonator. *FONETIK 2004 Proceedings.*

Gouvea. E., 1998. *Acoustic-Feature-Based Frequency Warping For Speaker Normalization*. Ph. D. Thesis, Department of Electrical and Computer Engineering Pittsburgh.

Đurić. N., D. Pekar, Lj. Jovanov, 2002. Structure of SpeechDat(E) database for Serbian, recorded over PTN. *DOGS 2002 Proceedings*, 1:57-60

Jakovljević. N., D. Pekar, 2005. Description of Training Procedure for AlfaNum CSR System. *EUROCON 2005 Proceedings.*

Pitz. M., 2005. *Investigation on Linear Transformations for Speaker Adaptation and Normalization*. Ph. D. Thesis University Aachen.

Uebel. L, P. Woodland, 1999. An Investigation into Vocal Tract Length Normalization. *EUROSPEECH99 Proceedings,* 6:2527-2530.

Welling L., H. Ney, 1998 Formant estimation for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6:36-48.

Zhan. P., M. Westphal, 1997. Speaker Normalization based on Frequency Warping. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing Proceedings*, 2:1039-1042.

Zhan. P., A. Waibel, 1997. Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition. *Language Technologies Institute Technical Report CMI-LTI-97-150*, Pittsburgh.