HIDDEN MARKOV MODEL-BASED APPROACH FOR NASALIZED VOWELS RECOGNITION IN SPONTANEOUS SPEECH

András Beke, Viktória Horváth

Research Institute for Linguistics of the Hungarian Academy of Sciences
beke.andras@nytud.mta.hu; horviki@ntyud.mta.hu

ABSTRACT

In this study, oral and nasalized vowels were analysed based on Hungarian spontaneous speech corpus. Although such results are generally based on analyses of isolated, read-aloud sentences, the authors suggested it is questionable that they are also true of spontaneous types of speech. There is a lack of agreement in the literature as to which measurable acoustic parameters correlate of nasality. MFCC as robust feature was presented earlier for nasalized vowel detection combined with SVM classifier. In this research, we investigated the use MFCC and HMM for automatic nasalized vowel recognition. Results support the view i) regressive nasalization could be classified with better accuracy than progressive nasalization, ii) the degree of nasalization strongly depends on the vowel quality, iii) low vowels show a large degree of articulatory nasalization, however, the acoustic consequences are smaller, therefore the perceived degree of nasalization is either similar or lesser than for higher vowels.

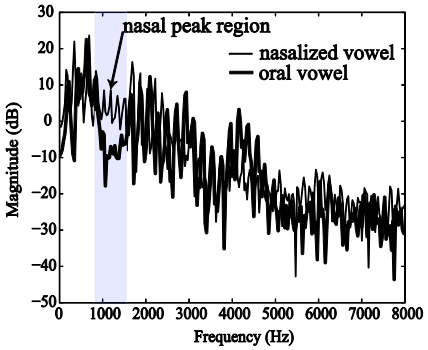
Keywords: vowel nasalization, spontaneous speech, automatic classification, Hidden Markov Models

1. INTRODUCTION

Vowel nasalization is the production of a vowel while the velum is lowered and the velopharyngeal port is open, so that the nasal cavity is coupled into the vocal-tract resonance system [11]. Nasalization can be described as an articulatory, aerodynamic, acoustic, and perceptual phenomenon as well.

A large number of studies have analysed the acoustic structures of nasalized vowels from several aspects in several languages (e.g., [1, 5, 12]). The degree of nasality of a vowel depends on the language, the vowel quality, the consonant quality [8], the phonological position of the syllable, and the speaker variability as well. The most important question of the researches is defining the most relevant acoustic parameters by which the nasalized vowels can be distinguished from the oral ones. Many acoustic parameters have been found to be related to nasalization: reduction in amplitude of the first formant (A1); the relationship between A1 and the amplitude of the first harmonic (H1); nasal poles, one of them below F1 (P0) at around (250–450 Hz), and the other one above F1 (P1) between A1 and P0, and the difference between A1 and P1, etc. [10] (Fig. 1).

Figure 1: Example for the nasal peak region in nasalized vowel.



The tongue position, the nasal airflow and the acoustic parameters of American English were analysed in nasal and in non-nasal context. The speakers raised their tongue during the articulation of [i]. Authors suggested that this tendency is a compensation for the low-frequency shift in spectral energy during the velopharyngeal opening. The lingual position of [a] didn't show any changes in nasal context. The effect of nasalization resulted in significant changes in COG in the vicinity of F1 in the case of [a], but not for [i] [2].

The articulation with electromagnetic articulography and the acoustic realization of three oral-nasal vowel pairs were analysed in French [3]. The F1 of nasal /e/ was higher and the F2 was lower than its oral pair's in all observed speakers' speech. The other vowel pair's production was not so solid in the speaker's speech.

There are some findings concerning the automatic classification of the oral and the nasalized vowels based on their acoustic features. Chen [4] extracted three parameters from the vowel's spectrum: A1, P1 and P0 values. The result showed that there was a good correlation between A1–P1 value and the degree of the nasality of the vowel. Hasegawa-Johnson et al. [7] applied a large set of APs (acoustic parameters) which included MFCCs and Support Vector Machine to distinguish nasal and non-nasal frames. The result

showed 62.9% accuracy on the test set. Pruthi [10] in his thesis used MFCCs and SVM to classify automatically the oral and nasal vowels in read speech. The result showed 79% accuracy on the test set. An MFCC and SVM system was built to distinguish between oral and nasalized vowels [15]. The result yielded 88.3% overall accuracy in nasalization detection. Oral and nasalized vowels from the TMIT database in clear and in noisy condition were classified based on MFCCs, A1–P1 using LDA and SVM based classifiers [9]. The SVM classifier yielded the best result in the detection of oral and nasalized vowels based on MFPSCC (69% on average in clear condition). The detection of /a/ was the most efficient (71%). SVM performed as a better classifier than LDA in detecting nasalized vowels in noise as well.

We suggest that nasalization cannot be described via only certain acoustic parameters. It is hard to identify the relevant formants or peaks. Another problem is staking out boundaries between nasal consonants and nasal vowels. LPC analysis cannot handle the anti-formants; and it is unknown whether the nasal effect increasing over time [12].

The previous studies on vowel nasalization were based on read speech corpus. Although such results are generally based on analyses of isolated, read-aloud sentences, the authors suggested it is questionable that they are also true of spontaneous types of speech.

In this study, we used data from a large Hungarian spontaneous speech database. MFCC was used in this study for the robust automatic classification of oral and nasalized vowels. In Hungarian, nasalization is not a phonologically distinctive mark, the vowels are nasalized due to nasal consonats' coarticulatory effect.

2. DATABASE

In this study, spontaneous speech (quasi monologue) of 19 native Hungarian speakers (10 males and 9 females; ages between 20 and 64 years) was used from BEA (BEszélt nyelvi Adatbázis 'spoken language database' in Hungarian, cf. [6]). In BEA, the spontaneous speech is recorded under silent chamber conditions using a microphone connected to a computer. Goldwave software is used to record the utterances. The sound files are saved in WAVE format at 44.1 kHz sampling rate and 16-bit PCM quantization. The phonetic transcriptions of all records were aligned with the speech waveform using Praat software for Speech Analysis. During the analysis, the authors used the following vowels: [ɔ], [a:], [ɛ], [i], [o] and following nasals [n] and [m]. Segmentations and alignments were carried out manually and controlled both visually and auditory. In the analysis, we processed 2,236 vowels in order to devise methods for the recognition of nasalized vowels.

3. METHOD

Mel-Frequency Cepstral Coefficients (MFCC) were calculated and used for the training of Hidden Markov Models (HTK implementation). The recognition system was trained on 1,490 vowels while testing was done on 745 further vowels.

3.1. Hidden Markov Models

In automatic speech recognition, Hidden Markov Models (HMM) are commonly used to model the phonemes of a language. In a speech recognition system, a dictionary specifies the pronunciation of words (dictionary entries) in the form of phoneme sequences, and a so-called language model specifies which word can follow a given word or word chain. The role of phoneme models is to map speech waveforms to phonemes. Mel Frequency Cepstral Coefficients (MFCC) used acoustic pre-processing method. The computation of MFC coefficients is as follows: first, a Fast Fourier Transformation (FFT) is applied to the speech waveform. Frequently, a 25-ms part of the speech sample is selected and weighted by a window-function (e.g., Hamming window). Then the window is shifted by the frame rate (usually 10 ms), and another FFT is done. In this way, a speech spectrum is obtained at every 10 ms. The second step of the pre-processing is the decomposition of the spectra corresponding to the critical bands of the human auditory system. This is done by a filterbank (e.g., a Mel filterbank) consisting of 20 separate band-pass filters. Each filter outputs the averaged energy in the given frequency domain covered by the filter. In this way, 20 values in each 10 ms can be obtained. The logarithm of these is taken and a Discrete Cosine Transform (DCT) is applied in order to de-correlate these values and reduce the dimensions to 12. This means that at this step 12 values–which form a vector or a so-called frame–represent each 10 ms of speech. Finally, by adding mean energy and calculating first and second order deltas, one obtains 39-dimensional feature vectors for each 10 ms.

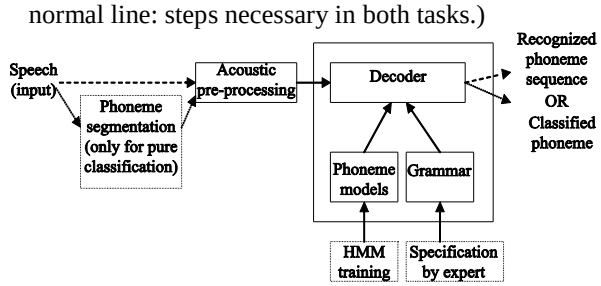
The phoneme HMMs model the distribution of the feature vectors that are assumed to be phoneme-specific. Phoneme HMMs are usually 3-state left-to-right HMMs in order to handle some coarticulation, too. Each state is assigned a probability density function, composed from a weighted mixture of

normal distributions (Gaussians) that characterize the "shape" of the feature vectors corresponding to the state.

During training, the parameters of these functions are estimated. When used for speech recognition, the feature vectors obtained by the same acoustic pre-processing are compared to the distributions estimated by the mixture. The more they fit, the higher the score of the actual state (sequence) will be when looking for the most probable hypothesis.

Indeed, HMMs in speech recognizers perform a classification task and an alignment task (they classify the phoneme realizations and detect their start and end points). The very same approach can be used to align a phoneme sequence to the input speech. In this case, phoneme classification and phoneme sequence alignment are performed in parallel, this is called phoneme recognition (Fig. 2). However, this approach can be further simplified by implementing a pure phoneme classification system where phoneme sequence alignment is not needed as each phoneme is pre-segmented and classified separately.

Figure 2: The integrated phoneme recognition or classification system. (Dashed line: steps needed exclusively for phoneme recognition; dotted line: steps needed exclusive for phoneme classification; normal line: steps necessary in both tasks.)



This task is called simply phoneme classification. Both for phoneme recognition and classification, phonemes and/or phoneme classes should be selected for modelling and then, for each class, the HMM should be trained using a statistically representative set of samples. Beyond the trained models, the recognition or classification task also needs a dictionary and a so-called grammar, which is a network or a finite state transducer composed from HMMs. In case of pure phoneme recognition/classification, the dictionary is not necessary and hence, the grammar specifies simply what kind of phoneme or phoneme-class sequences are allowed to be aligned to the input speech (phoneme recognition) or what are the classes used for the classification (phoneme classification) [15]. This system was implemented using the HTK toolkit [14].

3.2. Evaluation

There are three types of errors in recognition tasks: deletion, insertion, and substitution. A deletion error occurs if the recognizer misses a phoneme. (It does not identify it as a separate phoneme when aligning the phoneme sequence to the input speech. In classification, however, only substitution errors may occur.) If one discards deletion errors, a ratio is obtained which can be interpreted as classification performance; however, in this case the missed phonemes are excluded from evaluation, distorting the results compared to "classical" classification. In other words, "correct without deletion" rate is the classification rate of the identified phonemes.

4. RESULTS

A phoneme classification task was designed to analyse the separability of all full nasalized vowels merged "N" and all oral vowels merged "O". 3-state left-to-right models were trained using 2, 4, 8, 16, 32 Gaussians in output probability density functions. The grammar used for decoding allowed for both of "N" and "O" with equal weights (probabilities). The best classification result was yielded by the 16 Gaussian models. The results are shown in Table 1.

Table 1: Classification results for oral vowels and for nasalized vowels.

Category	Correct	Correct without deletion
N	72.84%	94.71%
O	77.36%	90.74%
ALL	75.82%	82.00%

Nasalized vowels were classified correctly in 72.84% of all nasalized vowel realizations.

The degree of vowel nasalization depends on vowel quality, especially height, therefore in the second experiment four HMM models were build: high oral vowels HO; low oral vowels LO; high nasalized vowels HN; low nasalized vowels LN. Typically, low vowels exhibit a large degree of articulatory nasalization (velopharyngeal port opening) but the acoustic consequences are smaller, so that the perceived degree of nasalization is either similar or lesser than for higher vowels.

The best recognition results for these four different vowels were obtained by 16 Gaussian models (see Table 2) in this phoneme recognition task.

The results showed that the recognition of low nasalized vowel was better than high nasalized vowel.

Table 2: Classification results for high/low oral vowels and for high/low nasalized vowels.

Category	Correct	Correct without deletion
HO	83%	90%
LO	71%	79%
HN	48%	59%
HL	77%	85%
ALL	73%	83.5%

In the next experiment, the progressive and the regressive nasalization effect size was testing. A phoneme classification task was designed to analyse the separability of all progressive nasalized vowels merged "PN" and all regressive nasalized vowels merged "RN". We supposed that the regressive nasalization has a greater effect size than progressive nasalization. The best recognition results for these four different vowels were obtained by 16 Gaussian models (see Table 3) in this phoneme recognition task.

Table 3: Classification results for progressive nasalized vowels and for regressive nasalized vowels.

Category	Correct	Correct without deletion
PN	56.47%	70.0%
RN	69.44%	90.20%
ALL	65.90%	82.25%

The results showed that the regressive nasalized vowel can be more precisely recognized than progressive nasalized vowel.

In the last examination, the HMM was built to recognize each nasalized vowel depend on vowel quality ([ɔn]:AN; [ɛn]:EN; [a:]:ÁN; [on]:ON; [in]/:IN). The best result was yielded using 16 Gaussian models (see Table 4) in this phoneme recognition task.

Table 4: Classification of nasalized vowels depend on vowel quality

Category	Correct	Correct without deletion
AN	77.77%	80.00%
EN	94.48%	95.23%
ÁN	85.00%	89.47%
ON	71.42%	76.92%
IN	57.14%	58.53%
ALL	82.00%	84.04%

The results showed that the classification of nasalized [ɛ] yielded the best result, the accuracy was 94.45%. The classification accuracy of [a:] (with the lowest tongue position in Hungarian) in nasal context was 85%. The classification of [i] yielded the poorest accuracy.

5. CONCLUSIONS

The aim of this study was the automatic classification of oral and coarticulatory nasalized vowels in Hungarian spontaneous speech.

Oral and nasalized vowels can be classified automatically with the accuracy of 75% with HMMs based on MFCCs. This method yielded better result than the SVM classifier for English oral and coarticulatory nasalized vowels [7], however, the methodological gap between automatic analysis of English speech produced under controlled conditions and different automatic analysis of Hungarian speech in spontaneous dialogues is wide to allow a proper comparison.

Different tendencies were found for the automatic classification of oral and nasalized vowels. The classification of higher oral vowels yielded better result than that of lower oral ones. However, the lower nasalized vowels can be classified with much better result than the high nasalized vowels.

The classification result of regression and progression nasalized vowels may lead to the conclusion that regressive nasalization has a greater effect on vowels than progressive nasalization. However, there is no general consensus in the literature that vowels may become more nasal when followed by a nasal consonant or preceded by a nasal consonant. This result also gives insight into possible differences in speech motor planning of carryover and anticipatory nasalization.

The degree of the nasal context depends on vowel quality. The acoustic parameters of Hungarian [ɛ] modified in the greatest extent based on the result of automatic classification. The result of classification was the poorest in the case of [i]. Other studies for English confirmed that the analysed acoustical parameters of nasalized [i] did not significantly differ compared to the oral [i], although the tongue was higher during articulation [2]. The raised tongue causes lowers F1, offsetting the acoustic effects of the nasal consonants. Speakers compensate the nasal effect on [i] but not on the other vowels in such a great extent. We suggest that there is less acoustic variability for [i] than [ɔ] on one hand. The velum needs to be more lowered for perception of nasalization for [ɔ] than for [i].

Based on Hungarian spontaneous speech data, we confirmed the results of previous research for other languages. However, we suggest that the acoustic analysis is not sufficient to clearly describe the effect of nasals on vowels' realization. Therefore, it needs to be combined with Electromagnetic Articulograph and/or nasometer as well.

6. ACKNOWLEDGEMENTS

This research is carried out with the support of OTKA 108762 project.

7. REFERENCES

[1] Beddor, P. S. 2007. Nasals and Nasalization: The Relation Between Segmental and Coarticulatory Timing. In: Trouvain, J., Barry, W. J. (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* Saarbrücken, 249–254.

[2] Carignan C., Shosted, R., Shih, C., Rong, P. 2011. Compensatory articulation in American English nasalized vowels. *Journal of Phonetics* 39, 668–682.

[3] Carignan C., 2014. An acoustic and articulatory examination of the "oral" in "nasal": The oral articulations of French nasal vowels are not arbitrary. *Journal of Phonetics* 46, 23–33.

[4] Chen, M.. Y. 1997. Acoustic correlates of English and French nasalized vowels. *J. Acoust. Soc. Am*. 102 (4), 2360–2370.

[5] Chen, N. F., Slifka, J. L., Stevens, K. N. 2007. Vowel Nasalization in American English: Acoustic Variability Due to Phonetic Context. In: Trouvain, J., Barry, W. J. (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* Saarbrücken, 905–908.

[6] Gósy, M. 2012. BEA – A multifunctional Hungarian spoken language database. *The Phonetician* 105/106, 51–62

[7] Hasegawa-Johnson et al. 2005. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop., Proceedings of CASSP 5, 213–216.

[8] Ladefoged, P. 2005. Vowels and consonants. Oxford: Blackwell Publishing.

[9] Najnin, S., Shahnaz, C. 2014. A detection and classification method for nasalized vowels in noise using product spectrum based cepstra. International Journal of Speech Technology. http://download.springer.com/static/pdf/997/art%253A10.1007%252Fs10772-014-9225-9.pdf?auth66=1421846665_c4d3ab786f69d2c6a9fe23a4d7a58334&ext=.pdf

[10] Pruthi, T. 2007. Analysis, vocal-tract modeling and automatic detection of vowel nasalization. http://www.isr.umd.edu/Labs/SCL/publications/theses/Pruthi_PhD.pdf

[11] Rose, P. 2002. Forensic speaker identification. London–New York: Taylor&Francis.

[12] Shosted R. 2012. A descriptive approach to the measurement of nasalization. http://faculty.las.illinois.edu/rshosted/docs/nwav41_nasal.pdf

[13] Vicsi K, Szaszák Gy 2010. Using prosody to improve automatic speech recognition Speech Communication 52:(5), 413-426.

[14] Young, S. J. – Evermann, G. – Gales, M. J. F. – Hain, T. – Kershaw, D. – Liu, X. – Moore, G. – Odell, J. – Ollason, D. – Povey, D. – Valtchev, V. –Woodland, P. C. 2006. The HTK Book (for HTK Version 3.4). University of Cambridge, http://htk.eng.cam.ac.uk, December 2006.

[15] Yuang, J. – Liberman, M. 2011. Automatic measurement and comparison of vowel nasalization across languages. Proceedings of ICPhS XVII, 2244–2247.