Adding Voicing Features Into Speech Recognition Based on HMM in Slovak

Juraj Kacur
Department of telecommunication
FEI, STU
Bratislava, Slovakia
kacur@ktl.elf.stuba.sk

Gregor Rozinaj
Department of telecommunication
FEI, STU
Bratislava, Slovakia

***Abstract*—This article discusses the impact of substituting some of the basic speech features with the voiced/ unvoiced information and possibly with the estimated pitch value.**

As a good measure of the signal's voicing the average magnitude difference function was assumed, especially the ratio of its average value to its local minima found within the accepted ranges of the pitch. Furthermore, the pitch itself was used as an auxiliary feature to the base MFCC and PLP features. Experiments were performed on the professional database SPEECHDAT-SK for mobile applications working in harsh conditions, using various HMM models of context dependent and independent phonemes. All models were trained following the MASPER training scheme.

In all cases the voicing feature brought improved results by more than 9% compared to the base systems. However the role of the pitch itself in the case of speaker independent ASR system evaluated over different tasks was not always so beneficial.

Keywords-speech recognition; speech features; HMM; AMDF MFC; PLP, MASPER

I. INTRODUCTION

There has been an immense research effort spent over several decades in order to realize any practical recognition system. Currently, automated dialog or even dictation systems are emerging in more or less limited way. However, there is still lot of to do to realize fully automatic transcription of natural conversation, as the practical systems should be robust, accurate, speaker independent, must support vocabulary sizes of several hundreds of thousand of words in the real time, etc.

These strict requirements can be met by the statistical speech modeling using HMM models of tied context dependent (CD) phonemes with multiple Gaussian mixtures [1]. Nowadays the classical concept has evolved into areas like: hybrid solutions with neural networks, different than ML or MAP training strategies, explicit time duration modeling, etc.

Beside the modeling problem, another vital issue of the recognition process is the feature extraction method (speech representation). This matter is not fully solved and various features have evolved during several decades, but the most successful ones in connection with HMM are MFCC and PLP. However, some other base features are also reported to be beneficial and are being able to outperform the previously mentioned ones, but these are usually closely attached to the certain environments.

Apart of the base static futures that aim to estimate magnitude-modified and frequency-warped spectra, dynamic features reflecting time evolution, like delta and acceleration coefficients proved to be rather valuable as well. As a consequence those static parameters eliminate the voicing information which carries some discrimination information, that play role in the recognition accuracy. Namely, this information is vital do discriminate some pairs of voiced and unvoiced consonants like: z in the word zero and s as appears in the word sympathy, p and b, d and t, etc. To elevate this drawback we substituted the least significant static features with such kind of information derived from the average magnitude difference function (AMDF). To test this concept further the pitch was included in separate experiments as well. In order to verify and asses the merit of such a modification, series of experiments were executed using the professional database and a training scheme for building robust HMM models for practical automatic speech recognition (ASR) systems.

The remaining article is organized as follows. First the base speech features represented by MFCC and PLP are introduced in brief. Then the process of extracting proper voice and pitch information based on the AMDF function is outlined, which will be followed by the detailed description of the testing conditions involved in the whole series of experiments ranging from the training database, training scheme and executed experiments. Finally, the article is concluded by commented results.

II. SPEECH FEATURE FOR ASR SYSTEMS

One of the first steps in any ASR system is to convert the incoming speech into proper features that would highlight the lexical information contained in it and suppress all adversary artifacts that would prevent us from estimating them. At the beginning it should be noted that this task is not yet completely solved and a lot of effort is still going on in this area. The aim is to simulate the auditory system of humans, mathematically describe it, and simplify for practical usage.

A good feature should be sensitive to differences in sounds that are perceived as different in humans and should be “deaf” to those which are unheeded by our auditory system. For example the location of formants in the spectra and their widths are important for sound discrimination. On the other hand, following aspects are not vital in perceiving differences: overall tilt of the spectra, frequencies located under the first and above the 3rd format frequency, narrow band stop filtering, etc.

Furthermore, features should be insensitive to additive and convolutional noises or at least they should represent them in such a way that these distortions are easy to locate and suppress in the feature space. Finally, when using continuous density HMM models it is required for the feasibility purposes that the elements of feature vectors should be linearly independent so that a single diagonal covariance matrix can be used.

Many basic speech features have been designed so far, but currently MFCC and PLP [2] are the most widely used in CDHMM ASR systems. They both represent some kind of cepstra and thus are better in dealing with convolutional noises. Furthermore, the DCT transform applied in the last step of the computation process minimize the correlation between elements and thus justifies the usage of diagonal covariance matrices. Besides those static features it was soon discovered that the changes in the time represented by delta and acceleration parameters play an important role in modelling the evolution of the speech, which is somehow restricted using the first order Markov chain. Finally, to take the full advantage of cepstral coefficients, usually a cepstral mean subtraction or temporal filtering using RASTA-style filters are applied in order to suppress possible distortions inflicted by various transmission channels or recording devices and promote speech components at the modulation frequency. In the following let us recap in brief the steps in calculating MFCC and PLP feature.

A. MFCC

The speech signal is first modified by HP so-called preemphasis filter to suppress the LP character of the speech given by the lip radiation to the open space. Prior to the FFT computation a Hamming window is applied and the frequency in Hz is warped into the Mel scale to mimic the critical bands over different frequencies. Next, equally spaced triangular windows with 50% overlap are applied to simulate a filter bank. Finally a logarithm is taken and the DCT transform is applied that produce a static feature frame. The logarithm not only acts as a tool to produce cepstrum (real one) but suppress the high-vale intensity in favor for low intensities as the human auditory system does. In addition, zero cepstral coefficient is used as well to estimate the overall log energy.

B. PLP

The original process of PLP calculation follows these steps: calculation of FFT that is proceeded by Hamming windowing, frequency warping into Bark scale, smoothing the bark-scaled frequency spectra by a window simulating critical bands effect of our auditory system, sampling the smoothed bark spectrum in approx. 1 bark intervals to simulate the filter bank, equal loudness weighting of the sampled frequencies which approximates the hearing sensitivity, transformation of energies into loudness by powering each frequency magnitude to 0.33, calculating the LP coefficients from the warped and modified spectra (all pole model of the speech production), finally cepstral LP coefficients are derived from LPC as if the logarithm was taken and an inverse FFT calculated.

To show some practical differences of these two features in the terms of recognition errors, in table 1 there are shown averaged WER for PLP and MFCC features as scored in the application words and digits string tests and the relative improvements achieved by subtracting zero mean, adding zero cepstral coefficient, delta and acceleration coefficients.

TABLE I. WER AND RELATIVE IMPROVEMENTS FOR MFCC, PLP AND THEIR AUXILIARY FEATURES AVERAGED OVER DIFERENT HMM FOR DIGIT STRINGS AND APPLICATION WORDS TESTS.

	Static WER [%]	**Relative Improvements to WER**			
		Zero mean subtraction [%]	***C0 [%]***	***Delta [%]***	***Acceleration [%]***
PLP	30.67	2.04	3.9	62.69	20.37
MFCC	33.12	5,99	15.8	50.56	13.61

As it can be seen the static features of PLP outperformed the MFCC counterparts by 7.9%. Adding another static coefficient C0 and applying zero mean subtraction is beneficial in all cases however it is more significant for MFCC. On the other hand, addition of dynamic coefficients was more relevant for PLP. These findings may suggest that PLP is bit better in describing static speech features for recognition purposes at least for the tested environment and settings that will be discussed later.

III. AMDF AND THE VOICING FEATURE

As it was already outlined the concept of incorporating some parameters assessing the voicing of a particular signal may bring additional discriminative information into the recognition process. For example in Slovak, one classification of consonants is according whether they exist in voiced / unvoiced pairs. In the paired group there are consonants grouped in pairs according to the mechanism how they are produced and perceived. In the paired group there is always an unvoiced consonant accompanying the voiced one. The only difference in their production is the absence of vocal chord activity that can not be observed after PLP or MFCC processing. Some typical pairs of voiced and unvoiced consonants are: p and b, d and t, k (king) and g (give), etc. As it can be seen, distinguishing between them may be crucial. On the other hand there are many cases (in Slovak) where in the real conversation the voiced one takes the form of unvoiced consonant and vice versa.

There are more methods to asses the volume of voicing and to detect the periodicity ranging from simple algorithms in the time domain like AMDF, and autocorrelation, through spectral ones like harmonic spectra and real cepstrum to methods operating on the residual signal after the inverse filtering. The advantage and the reason why we opted for AMDF is a simple and fast implementation and good results even in lower SNRs. AMDF is defined as follows:

$$f_i(k) = \sum_{n=0}^{n<N} \left| s(i \cdot N + n) - s(i \cdot N + n + k) \right| , \; k \in <Tmin,Tmax>, \qquad (1)$$

where s is a signal, N is block's length and Tmin and Tmax are the minimal and maximal pitch periods. Then as a measure of voicing the minimal value of AMDF can be taken. To suppress its dependence upon magnifying constant, usually a ratio to its maximal or averaged value is computed as follows:

$$\text{voicing}_i = \frac{\dfrac{1}{T_{\max} - T_{\min}} \sum_{k = T_{\min}}^{k < T_{\max}} f_i(k)}{\min\limits_{T_{\min} \le k < T_{\max}} \left(f_i(k) \right)} . \qquad (2)$$

Thus perfectly periodic signals would produce infinity, whereas signals with no period would be close to 1, but still higher. Then the location of the minima can represent the estimated pitch; however certain precaution should be taken not to detect longer periods- usually integer multipliers of the real one.

IV. Experiment Setting

In this paragraph basic settings of the experiments will be given for clarity. These are related to the used database, training process and models, and finally to the evaluation tests.

A. *Speech database: Mobildat-sk*

As both the training and recognition tasks are more challenging in the adverse environments the Slovak MOBILDAT database [3] was chosen. It was recorded over GSM networks and generally provides more difficult conditions.

The MOBILDAT-SK database consists of 1100 speakers that are divided into the training set (880) and the testing set (220). Each speaker produced 50 recordings in a session with the total duration ranging between 4 to 8 minutes. These items were categorized into the following groups: isolated digit items, digit strings, yes/no questions, dates, times, application keywords, directory names, spellings, phonetically rich words, and phonetically rich sentences. Description files do not contain any time marks and beside the speech several non – speech events are labeled e.g.: truncated recordings, unintelligible speech, filed pauses, speaker noise, GSM specific distortion, etc. In total there are 15942 different Slovak words, 260287 physical occurrences of words. Finally, there are 41739 useable speech recordings in the training portion, containing 51 Slovak phonemes, 10567 different CD phonemes (word internal) and in total there are slightly more than 88 hours of speech.

B. *Training process of HMMs*

The training process is based on MASPER [4] training scheme designed for building multilingual and cross lingual reference recognition systems. As its thorough description can be found elsewhere, here only the main features are listed. All phonemes are modeled with 3 state models following the Bakis structure and 4 non speech events are modeled too: short pause model (one state T model), background model (3 state, backward connection allowed), speaker generated noise and a hesitation model (both are 3 states with the Bakis structure). Recordings which contain damaged speech like: truncated, mispronounced or unintelligible words are completely removed from the training. There are 3 runs of the training, where the first two produce CI phonemes with 1 up to 32 mixtures. The first one uses a flat start initialization and an embedded training, and the second run of initialization and several cycles of the single model training operates on the time aligned recordings (multiple pronunciations are aligned too) utilizing the Viterbi training and CI models from the first stage. These are only applied to single mixture CI models and by doing so more accurate models are obtained than their counterparts from the first run. Next the single mixture CI phonemes from 2nd run are used to clone CD phonemes, which are further tied by the decision trees that were constructed using the phoneme classification file designed for Slovak language. Finally, tied CD phonemes are trained in cycles with gradually incrementing the number of mixtures from one to 32. In the following experiments both PLP and MFCC basic vectors with 12 coefficients plus the C0 were used together with the delta and acceleration parameters (all together 39 elements). When the voicing ratio or pitch were used, they substituted the least significant basic features, i.e. they replaced 12th and 11th elements of PLP and MFCC vectors respectively as we believed the carry the least discrimination information (finest details of modified spectra). This was vital to make the comparison and assessment as objective as possible, so final vectors in all tests were of the same length of 39 elements.

C. *Evaluation tests*

Two basic tests were executed upon all models and the examined feature extraction methods, namely: MFCC and PLP both in combination with the voicing and pitch estimates. All experiments were accomplished on the test part of the MOBILDAT-SK database that amounts to 220 speakers. The two basic tests were: digits strings that can contain arbitrary number of digits in the string and the application words test that equals to the recognition of isolated words. Although the digits string test exhibits higher perplexity and therefore provides higher errors it uses only very limited set of CD phonemes. On the other hand, the application words test contains greater variety of words, CI and CD phonemes and therefore provided more objective insight about how good all the models were trained.

V. Results

To asses the merits of incorporating voicing measure and the pitch into the recognition process, achieved results are related to the original ones (MFCC and PLP) via the relative improvement unit defined as follows:

$$\text{relative\_improvement} = \frac{WER_{orig} - WER_{mod}}{WER_{orig}} 100 \; . \qquad (3)$$

Further it should be noted that the new features were not simply added to the original vector but instead they replaced its least significant elements (we believed according to the theory they were the last PLP and MFCC coefficients). This allowed us to maintain the same vector size and made the comparison more objective.

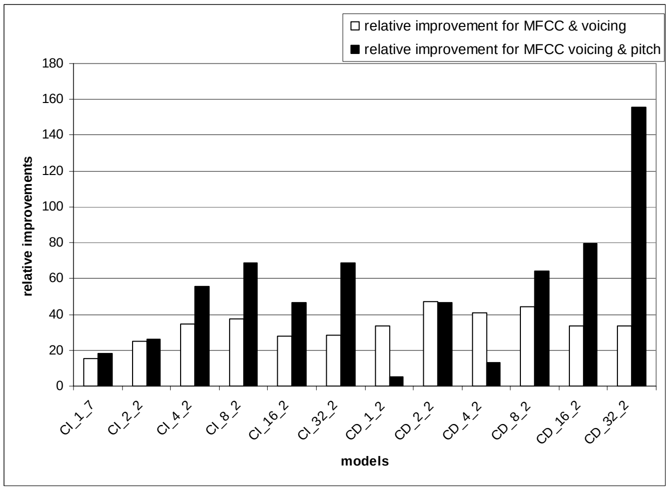

Figure 1. Relative improvements for MFCC, various HMM models (CI and CD with 1 to 32 mixtures each) and application words test achieved by incorporating voicing and pitch measures.

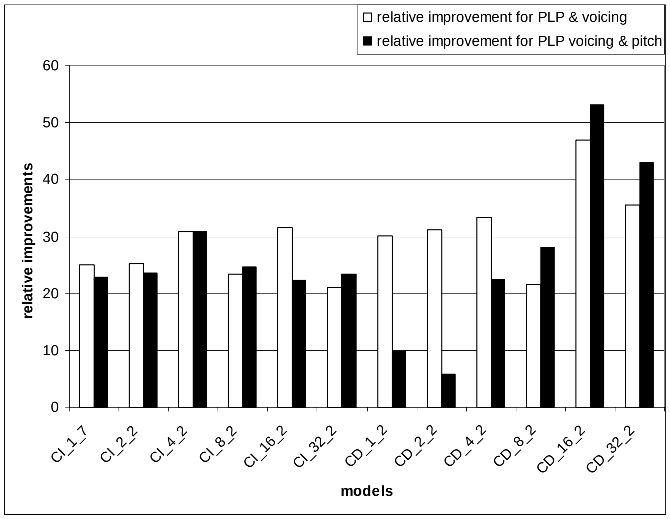

Figure 2. Relative improvements for PLP, various HMM models (CI and CD with 1 to 32 mixtures each) and application words test achieved by incorporating voicing and pitch measures.

As the PLP and MFCC features performed in a slightly different way the results will be given separately for both features. Finally, to save up the space only results for the application words are further showed, as similar findings were observed in the case of digits strings, however due to the higher perplexity lower accuracies were achieved.

In fig.1 there are depicted the relative improvements for MFCC achieved by introducing the voicing measure and the pitch feature in the case of application words test. The same results for PLP are shown in fig.2. In all the cases it is clear that better results were observed by the inclusion of suggested features, however the replacement of the least significant basic features is more relevant for MFCC where on average an 32% improvement was recorded whereas for PLP it was only 17%. Further, it is noticeable that the pitch was much more successful (black bars) in connection with MFCC features and that its incorporation tended to be more beneficial in the case of more complex models (with higher number of mixtures) for both PLP and MFCC parameters.

VI. CONCLUSIONS

Incorporating the voicing coefficient defined in equation 2 and 3 that assess the level of similarity between possible periods of voiced signals turned out to be very effective, on average a 24.5% improvement was observed for MFCC and 19.9% for PLP. This feature may provide necessary discriminative information between paired consonants (voiced and unvoiced, like: s and z, p and b, t and d, etc.).

The inclusion of pitch itself by replacing the least significant PLP or MFCC elements was despite its dispersion beneficial in case of MFFC, where the improvement on average reached 32.9%, which means a further 9% improvement comparing to voicing alone. In the case of PLP opposite situation was recorded, the improvement on average reached only 17.6% which is actually a reduction by more than 2% comparing it to the test with voicing feature alone. On the other hand the pitch was successful for more complex models even in the PLP case; however benefits were relatively minor comparing it to MFCC. This phenomenon can be explained by the better description capabilities of more complex models that are more equipped to cope with the wider spectra of possible pitch values.

These results with voicing and pitch parameters are in line with those mentioned in table 1, where it was shown that PLP static features were better in representing the speech for recognition purposes in the tested settings. Here the replacement of the least significant features by voicing and pitch parameters was more relevant for MFCC extraction method in all cases.

ACKNOWLEDGMENT

The authors hereby declare that this research was founded by the grant: VEGA – 1/0718/09.

REFERENCES

[1] X. Huang, Y. Ariki and M. Jack, “Hidden Markov Models for Speech Recognition”, Edinburg university press, 1990

[2] F. Hönig, G. Stemmer, Ch. Hacker, and F. Brugnara, “Revising Perceptual linear Prediction (PLP)”, Proceedings of INTERSPEECH 2005, pp. 2997-3000, Lisbon, Portugal, Sept., 2005

[3] S. Darjaa, M. Rusko and M. Trnka, “MobilDat-SK - a Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak”, In Proceedings of SPECOM'2006, Anatolya publishers, St. Petersburg, 2006, pp. 449-454, ISBN 5-7452-0074-x

[4] B. Lindberg, F. T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kacic, A. Zgang, K. Elenius and G. Salvi, “A Noise Robust Multilingual Reference Recognizer Based on SpeechDat(II)”, In Proceedings of ICSLP 2000, Beijing, China, October 2000