

An Assessment of the PhD. Thesis
“Automatic Speech Recognition of Vietnamese”
written by **Nguyen Thien Chuong**

Assessor: Prof. Ing. Václav Matoušek, CSc., ZČU Plzeň

Mr. Thien Chuong introduces in his thesis comprehensive problems of automatic speech recognition of Vietnamese. The thesis seems me as the basic work introducing all necessary problems to be solved at the development of the speech recognition system – the selection of phonetic units to build acoustic models, the collection of text and speech corpora, the creation of pronouncing dictionary, the construction of language model and the methods to deal with tone. Several methods for collecting large text and speech corpora are described in his work, two types of text corpora were obtained by exploiting the source of linguistic data from the Internet, and two types of speech corpora were extracted, including an Internet-based large continuous speech corpus and recorded audio-visual speech corpus.

Based on these methods and corpora a standard phoneme set optimal to Vietnamese with his corresponding grapheme-to-phoneme mapping table was proposed, various types of pronunciation dictionaries and language models for Vietnamese were constructed, and the optimal way to integrate tone in a syllable as well as the strategies to deal with speech recognition of Vietnamese were examined in the form of large vocabulary continuous speech recognition tasks. Finally, many types of visual front-ends and visual features were examined in the task of isolated word speech recognition of Vietnamese.

The presented topic can be seen as modern and clearly up-to-date from the point of view of the subject field development state – automatic recognition of Vietnamese speech. The author's approach to the topic was very responsible, he perused a lot of relevant works, he performed well-founded analysis of the present state of problems being solved in the work, and he made a creditable piece of work illustrated by several interesting examples. The most important results of his work are summarized and presented in the conclusion (last chapter) in order to connect the contributions together and build a bigger and more thorough picture of the presented problem.

The articulation of the doctoral thesis into chapters (nine chapters of the text and very interesting appendices) is logical; the motivation of his work is given in the first chapter and the goals of the doctoral thesis (major contributions of this work) are given in the fourth chapter. The second and third chapters give a nice overview of the field of automatic speech recognition and possibilities and state-of-art of the Vietnamese language and speech recognition. The significant results of the work are presented in chapters 5 to 8, each chapter contains a closed part of the topic. Chapter 9 (conclusion) summarizes then seven main tasks which have been proposed, described and experimentally evaluated in order to get a practically applicable ASR system for Vietnamese. The significant results of the thesis are illustrated by the special experiments and tests whose results are stored on the enclosed CD.

Summarizing the asset of the thesis, it offers the collection of three types of text and speech corpora which were used for experiments on speech recognition tasks, the very difficult and also the most interesting problem in LVCSR of Vietnamese – modelling tone in

syllable, definition of standard phoneme set for Vietnamese, audio speech recognition of Vietnamese, and audio-visual speech recognition.

The doctoral thesis written by Mr. Thien Chuong is evaluated positively; I have no doubt about the values of the topic and results of the work. The used methods, standpoints and the verification manners and presentations represent not only the valuable scientific results, but they also represent that the author is the forward-looking worker with corresponding erudition. The content of the presented doctoral thesis will certainly be used as the good source material for many other doctorate students and researchers.

I have only few comments and queries to the content of the thesis; they have only the formal character:

- Why is incoherence one of the prominent characteristic of works related to ASR of Vietnamese ? Only of Vietnamese ?
- Why did you used three pronunciation dictionaries (p. 52) ?
- You described a syllable-based strategy very simple (9 lines). Why (p. 64) ?
- Why the scheme on Fig.7.1 ends by the perplexity (p. 77) ?
- Could you a bit precise your assertion "... inner frame and across frames LDA will be trained. For inner frame LDA, visual feature vector extracted from each video frame will be considered as a sample to train LDA matrix. Using this type of LDA features, we can select not only the first few highest energy visual coefficients but also some other types of coefficients which capture more useful information from ROI of mouth." (p. 101) ? What other types of coefficients ?
- Did you use the "isolated word visual only speech recognition" only for comparison with the audio-visual speech recognition ?

The content of the thesis positively supports the competence of the author to apply and successfully implement the elected theoretical resources. The illustration examples were chosen appropriately and follow up the theoretical conclusions very well. From the formal point of view, the doctoral thesis is written with several printing errors, but it is of an appropriate graphical level.

Conclusion:

I have reached the conclusion that this work brings several new pieces of knowledge; the core of this work was correspondingly published. The work can be qualified as a good doctoral thesis and I am happy to recommend to the Scientific Council of the Faculty of Mechatronics and Interdisciplinary Engineering Studies of the Technical University of Liberec to award to the candidate the degree of doctor in philosophy.

In Pilsen on 23. November 2014



Prof. Ing. Václav Matoušek, CSc.

Department of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia in Pilsen

REPORT ON DOCTORAL THESIS

Title of the thesis: Automatic Speech Recognition of Vietnamese

Ph.D. candidate: Ing. Nguyen Thien Chuong

Reviewer: Doc. Ing. Petr Pollák, CSc.
Czech Technical University in Prague,
Technická 2, 166 27 Praha 6, Czech Republic

The work of Nguyen Thien Chuong deals with the automatic speech recognition (ASR) focused on Vietnamese language. The importance of this research is evident. Nowadays, the achieved accuracy of ASR systems is generally very high for major word languages and the research on a language where ASR technology is not well developed yet or when the language is rather under-resourced is really the important task for the speech research community over the world. The importance of the improvement of the state of the art on Vietnamese ASR is also evident due to the number of native speakers (Vietnamese is among the top 20 languages over the world from this point of view), moreover, it could be also very interesting in our country due to the significant Vietnamese community living in the Czech Republic.

After standard introductory chapters, the author presents the goals which can be accepted as goals of the dissertation thesis. It can be said that they were accomplished, however, I also see several issues spoiling a little other good results of this work. The most important contributions, several remarks and questions for the discussion during the forthcoming defence are summarized in the following points.

- The first general contribution is in the overview about Vietnamese, i.e. the language itself, its phonology, and further specific features important for speech recognition purposes. Such overview citing the most important current international research is important and useful, especially, when Vietnamese is the language for which the research on ASR is not so world-known.

On the other hand, some parts of this overview are too long, redundant, and not significant for the further research on Vietnamese (e.g. some parts of 2.1 or in many details described general history of ASR in section 2.2). I also missed the more detailed comparison of Vietnamese to other languages of the same origin. The author mentions differences between Vietnamese and Western languages as English or French, but why experiences with Chinese or other tonal Asian languages are not presented in more detail?

- Secondly, the author introduced the proposition of strategies for Vietnamese speech recognition, i.e. the phoneme set proposal, the grapheme-based or phone-based dictionary creation, the ASR approaches containing the tone modelling suitable for the Vietnamese which are based on phoneme-based, vowel-based, rhyme-based, or syllable-based strategies. These approaches of tone incorporation into Vietnamese ASR comprise the first principal contribution of the thesis.

However, the tone incorporation is done purely at the model level with standard MFCC feature set which does not contain any information about the tone (or prosody in general). The feature level incorporation is mentioned only briefly in the 2nd paragraph at page 41 and it is then not used anymore within experiments. It should have been included at least in a discussion about it because other authors use pitch information (sometimes also other prosody features) for a tonal language ASR various ways.

Concerning features, the experiment described in section 7.2 seems to be really redundant

for PhD thesis. The predictor coefficients are clearly features which are not suitable for the recognition and they are generally not used in ASR systems.

- Thirdly, the language modelling for LVCSR was studied. This part seems to be the second dominant contribution of this work and the author has paid the highest attention to this topic in the report in the parts describing the created corpus, the LM creation and testing at the text level, and finally also its impact in target LVCSR.

Here, the general note about the usage of HTK toolkit in all realized LVCSR tasks is mentioned. Was it also the case of the used decoder? If so, which one and what about your experiences from the point of view processing time? Sometimes it is difficult to reach real-time processing with HTK tools, especially for complex LVCSR.

Concerning the results of experiments, no numerical results of other authors are mentioned, neither within the state of the art nor when the own achieved results are presented. Many authors has achieved worse results, on the other hand, authors of [114] presents SER about 10%. I understand that an exact comparison is not possible due to possible differences in the setup of experiments or due to various ways of classification (WER vs. SER). Nevertheless, more precise discussion taking into account other achieved results could be included.

- Fourthly, the created text and speech corpora for Vietnamese ASR comprise further significant result of this work. The text corpus containing 8M sentences with near 200M syllables and speech corpus of approximately 50 hours of speech from 196 speakers could well contribute to further research in ASR of Vietnamese. This represents very important result for the realization of speech recognition for any language.

On the other hand, some important information about the collected databases is missing. What about sampling frequency of recordings in the collected speech corpus, the statistics about the coverage of particular phones or graphemes for the training, etc.? Moreover, in section 7.3.4 another collection of speech data for gender recognizer is described and again with only brief description. This is finally a little bit confusing within the reading the thesis. The author also describes in details several other databases of Vietnamese in section 3.3, but they are not used in this work. Why some of these databases were not used? E.g. GlobalPhone databases are available for a reasonable price.

Finally, it could be interesting for the research community, how it is with the availability of created databases. Will these databases be available for the research community either freely or commercially via ELRA or LDC?

- In the end, the author performed the study which has shown the impact of audio-visual recognition in the case of noisy speech recognition for Vietnamese. For this purpose the audio-visual speech corpus for Vietnamese was created.

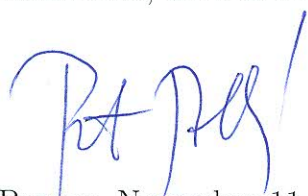
I see this topic slightly redundant for the work with the title "Automatic speech recognition of Vietnamese". The author shows the results confirming the impact of visual information for noisy speech recognition and mentions the proposition of method CLM (at the top of page 99), however, CLM is also mentioned as known technique at the previous page. What is the difference? Also no newly proposed technique is mentioned in the conclusion. Finally, if I understand it well no specific processing from the point of view Vietnamese was used within audio-visual speech recognition (excepting the above mentioned Vietnamese corpus created). From this point of view I feel it as a topic for another work.

Concerning the formal issues, the dividing into particular sections is sometimes not well chosen, e.g. there are 3 subsections in the Introduction for several paragraph, on the other hand, there is very long section 3.3 (17 pages of State of the art ASR of Vietnamese) where the reader has worse orientation within various problems described in this block. Concerning English, as a non-native speaker, I will not assess in details the level of English. I think that it is generally good, I have not found many errors. Concerning the bibliography, if the list of cited works is rather huge like here, the alphabetical ordering could be better to join the works of same authors.

The author published the results of his research within 5 works published at international conferences and workshops as: TSD, IDAACS, TSP, ECMMS. Although the publications at prestigious international conference (e.g. Interspeech or ICASSP) or journal articles are missing, published works are indexed in Scopus database and they were published at events with papers available via databases as Springerlink or IEEE Explore which proves the acceptable originality of presented results at international level.

Overall, the thesis of Nguyen Thien Chuong shows his capability of independent and original research activity. The remarks mentioned above do not doubt this fact, the thesis presents significant research work done over his Ph.D. study and it should contribute the progress in the research field on Vietnamese speech recognition.

On the basis of above mentioned facts, **I do recommend** the thesis for the presentation with the aim of receiving the Doctoral degree at Technical University of Liberec, Faculty of Mechatronics, Informatics and Interdisciplinary Studies.



In Prague, November 11, 2014