Switchboard Workshop 1998 の経過とその成果

シュスター・マイク
ATR 音声翻訳通信研究所
〒 619-02 京都府相楽郡精華町光台 2-2
Tel.: 0774-95-1394 e-mail: gustl@itl.atr.co.jp

あらまし
本稿では １９９８年度に合州国ジョンホプキンス大で行われた音声認識ワークショップの経過とその成果について報告する。このワークショップの目的は世界各地の研究者に６～８週間の協力を仰ぎ、音声言語処理のいくつかのトピックに取り組むことにある。本稿の第一章ではワークショップの構成に関して、第二章では本年度に取り上げられた三つのトピックに関して簡単に説明する。詳細に関しては、本ワークショップのメンバーの一部で構成される "Dynamic Segmental Models for Speech Articulation" グループから報告がなされている。

キーワード • 音声認識のワークショップ

Switchboard Workshop 1998 – Impressions and Results

Mike Schuster
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02
Tel. 0774-95-1394 e-mail: gustl@itl.atr.co.jp

Abstract
This paper describes experiences and results of the 1998 speech recognition workshop held yearly at Johns Hopkins University in the U.S. The aim of the workshop is to bring together researchers from all over the world to work intensively on a few selected topics within speech and language processing for 6-8 weeks. This paper describes in the first part how the workshop is organized. In the second part the three topics of this year's workshop are briefly introduced. More detailed information is given for the "Dynamic Segmental Models for Speech Articulation" group, since the author was part of it.

key words • speech recognition workshop

1 INTRODUCTION

Speech recognition is a very interesting and difficult topic, and it has currently a very active research community all over the world. Since around 1992 it became more and more popular to measure the performance of speech recognition systems and their algorithms by the use of standardized tests, which avoids many difficulties in comparing systems that are built using different training and test data.

The joint effort of the research community is aimed at improving current basic speech recognition technology. For the experiments to prove new ideas this requires in many cases a state-of-the-art speech recognition system, which consists in general of a complex collection of difficult algorithms and implementation details. Building a state-of-the-art system from scratch needs 1) rare experts which fully understand speech recognition theory and have practical experience to know what modules are necessary to build a state-of-the-art system, and 2) experts which are able to implement and use the algorithms and ideas efficiently and correctly, such that large scale experiments can be run. Today's systems are so complex, that an individual or even a small group of people cannot be expected to improve current speech recognition technology by a significant amount. To stimulate joint research across groups and countries, several years ago a series of yearly workshops was initiated by Prof. Fred Jelinek at the Center for Speech and Language Processing (CLSP), Johns Hopkins University in Baltimore, Maryland, U.S.

This paper summarizes the impressions and results of the sixth workshop 1998 from the view of the author, who was part of one of the groups. In the first part it is described how the workshop is organized, what its goals are and how it is funded. In the second part the three topics of this year's workshop are briefly introduced. More detailed information is given for the "Dynamic Segmental Models for Speech Articulation" group, since the author was part of it.

2 THE WORKSHOP IN GENERAL

2.1 History

The workshop was held every year since 1992, making the 1998 workshop the sixth in the series.

2.2 Goals

The goal of the workshop is to bring together top-level researchers from all over the world to work on specific topics of speech and language research for six to eight weeks and to make the results publically available on the WWW at:

`http://www.clsp.jhu.edu/ws98/`

2.3 Organization

The workshop is organized by Prof. Fred Jelinek (former leader of the IBM speech research group) and his team. The team consists of several researchers, two organizers/secretaries, one system administrator and several part time students, which were hired specifically for the workshop.

2.4 Funding

The first five workshops were funded solely by the Department of Defense (DoD). The 1998 workshop was co-funded by the National Science Foundation (NSF).

2.5 Topic selection and team preparation

The topic selection and team preparation was done as:

- The CLSP team invites about thirty researchers who are believed to be important in the field for a meeting held over a few days in October/November in the year before the workshop is going to be held.
- Potential team leaders propose a research topic that fits into a time frame of six weeks and that can be worked on with several researchers. One condition for all workshop participants including team leaders is that they have to be continuously present for the duration of the workshop, which rules out some members of industrial research labs as leaders.
- Three or four teams are selected for participation in the workshop in July/August of the following year.
- The complete teams meet once or twice *before* the workshop actually starts to get to know each other, to exactly explain the project and to prepare software and other necessary tools.

2.6 People

The workshop has about 30 participants – in 1998 there were 17 researchers, 7 graduate students and 6 undergraduate students. Participitation is based on invitation.

2.7 Duration

The workshop is held for six weeks in the summer, with two optional weeks preparation time for undergraduate and graduate students before the actual workshop starts. In the two preparation weeks this year there were introduction lectures about speech recognition (Fred Jelinek), machine translation (Fred Jelinek), information theory (Sanjeev Kuhndanpur), signal processing (Joe Picone) and HTK (Bill Byrne). For some lectures there were exercises for the undergraduates to get used to to the topics.

2.8 Location

All recent workshops were held at the Center of Speech and Language Processing (CLSP) at the Johns Hopkins University in Baltimore, Maryland, U.S. Everybody works in one big room, such that interaction between researchers is easily possible.

2.9 Housing

All participants get a room or an apartment, some people brought their families, too. All housing is within walking distance from CLSP.

2.10 Equipment

CLSP provides a number of desk-top machines and running machines for each team, such that large scale experiments are possible.

2.11 Communication

One of the most important features of the workshop is to stimulate communication between the researchers. This year there were:

external (invited guest) lectures: Five or six invited guest speakers (one a week) give one-hour lectures about their research. Detailed information is given below and can also be found at the CLSP WWW site.

internal lectures: People attending the workshop give lectures about their research outside of the workshop. All talks can be found at the CLSP WWW site.

team updates: Team members give twice during the workshop an update about the team's progress, which are also documented at the WWW site.

other social activities: There were several other social activities to enhance communication between people:

- every day at 3 pm cake and coffee for everybody
- once a week group dinner
- barbeque at start and end of workshop

2.12 Final presentations

The last two days of the workshop are reserved for the extended final presentations of the teams. Used methods, results of experiments and general experiences are reported. They are attended by the sponsors and many university and industry researchers.

3 SWITCHBOARD CORPUS

Since most of the experiments are done using the Switchboard corpus, which is currently one of the most interesting databases for speech research, a short description is given here.

The Switchboard corpus is a 120 - 160 h spontaneous speech corpus of telephone speech, which is distributed by the LDC.

`www.ldc.upenn.edu`

It contains conversations between people that talk about a given subject. The first transcriptions contained about 10% word error rate, meaning that on 10% of the data the human transcribers couldn't agree on a unique transcription, showing how difficult it is even for humans to recognize this data. The Switchboard corpus contains all kinds of telephonic spontaneous speech phenomena, like hesitations, false starts, background noise, low signal to noise ratios, unknown words, etc. There are several test sets – the 1997 testset has about 2200 utterances. Because it is spontaneous speech, the primary factor for good results is the quality of the acoustic models.

The Switchboard corpus is used by many groups to test their state-of-the-art speech recognition systems. The Switchboard corpus is also used for one part of the yearly evaluations of speech recognition systems by NIST:

`www.itl.nist.gov/div894/894.01/test.htm`

A description of all the systems that took part in the 1998 evaluations can be found in the `tar` files at the ftp server:

`ftp://jaguar.ncsl.nist.gov/`

in the directory

`evaluations/hub5/sept98/results/`

Currently, the best error rates in the evaluations are around 40% on the 1998 test set, using all known techniques to recognize speech. Decoding time and memory usage for these state-of-the-art systems is about 200 times realtime with 500 Mb of memory. The large error rates make the Switchboard corpus an excellent database for research in acoustic modeling.

4 THE 1998 WORKSHOP

This year's workshop was held from July 13th to August 21st, 1998, at the CLSP. There were three teams:

- Dynamic segmental models for speech articulation
- Rapid speech recognizier adaptation for new speakers
- Core natural language processing technology applicable for multiple languages

4.1 Dynamic segmental models for speech articulation

The leaders of the first group were John Bridle (Dragon Systems, UK) and Prof. Li Deng (University of Waterloo, Canada). All information about the group and details about their work can be found at:

`www.clsp.jhu.edu/ws98/projects/dynamic`

We tried to use a new type of model (Li Deng's DSM and John Bridle's R&B model) for speech recognition. The general idea for using this type of model is to introduce more prior knowledge about the state progression of natural speech [1, 3, 2].

As training data we used only about 1 h of Switchboard data from a single speaker, because of training problems the model couldn't deal with multiple speakers yet. Testing was done on 1241 VTLN warped utterances from male speakers from the 1997 Switchboard test set, which had 23 speakers.

Because it was yet unclear how to use the models in a Viterbi search, we used them to do only rescoring of N-best lists by comparing acoustic scores, which came from another, regular HMM system. The main result was [2], that the new models where better at picking the reference transcription than any other model, when exposed to the 5-best transcriptions plus the correct reference. Unfortunately they were not better than the other tested models (see Tab. 4.1), when the reference was taken out. The results are still preliminary and more research has to be done to see whether these models can improve current speech recognition technology or not.

MODEL	Ref+5	5-best	100-best
best possible	0.0	42.7	60.2
chance	45.0	54.0	60.2
baseline HMM	48.2	52.0	56.9
best R & B	32.7	52.6	59.4
best DSM	32.2	54.5	59.7

Tab. 1: Word error rate for an N-best rescoring task. The new models outperform the traditional models, when the reference is included into the N-best list for N=5.

4.2 Rapid speech recognizer adaptation for new speakers

The leader of the second group was Vassilis Digalakis (University of Creete, Greece). All information about the group and details about their work can be found at:

`www.clsp.jhu.edu/ws98/projects/adapt`

The aim of this group was to test *unsupervised* and *supervised* rapid adaptation techniques for speakers (same phone call within the Switchboard corpus). It is beyond the scope of this paper to describe the used techniques in detail – please refer to the WWW site above. Here just a brief discussion with the main results is given.

Used techniques were MLLR, multiscale trees, strutured MAP, Markov Random Fields and speaker normalization with the all-pass transform. As data the 1997 Switchboard testset was used (about 2200 utterances). Adaptation was run as

I) *batch mode adaptation*, meaning to adapt on the first 30/60 seconds of the conversation with the correct transcriptions (supervised), or with the transcriptions from the recognizer

(unsupervised), and test on the full second half, or

II) *transcription mode adaptation*, meaning to adapt and test on the output of the recognizer (unsupervised).

Baseline results on the speaker independent testset were 45.3% word error rate. The best results after adaptation were 40.2% error for supervised adaptation and 42.0% error for unsupervised adaptation, using 60 seconds of adaptation data.

4.3 Core natural language processing technology applicable for multiple languages

The leader of the third group was Prof. Jan Hajic (Charles University, Czechia). All information about the group and details about their work can be found at:

`www.clsp.jhu.edu/ws98/projects/nlp`

The aim of this group was to apply and improve several parsers to Czech to generate parse trees. The group started out with two statistical parsers and three rule based parsers. Besides improving the original parsers one of the efforts of the group was to combine all parsers into one.

The results went during the workshop from the baseline (72.3%) parse accuracy to 80.0% using the best statistical parser.

4.4 Invited guest lectures

In 1998 there were five guest lectures. All used slides are available on the WWW at:

`www.clsp.jhu.edu/ws98/schedules/guest.html`

The lecturers and there talks were:

- Lauri Karttunen (Xerox Palo Alto Research Center), "Core natural language processing technology applicable to multiple languages",
- Chin Lee, (Lucent Technologies), "A detection approach to speech recognition and understanding"
- Tony Bell, (Interval Research Laboratories, Palo Alto, CA), "Source separation and learning non-orthogonal bases for signals using Independent Component Analysis"
- John Lafferty, (Carnegie Mellon University), "Probabilistic models for clustering natural language data"
- Mark Gales, (IBM, T.J. Watson Research Laboratory), "Constrained estimation of Hidden Markov Models"

5 CONCLUSIONS

The latest results on the spontaneous speech evaluation tests show that spontaneous speech recognition is still far from being a solved problem. The best state-of-the-art systems built by experienced groups with years of man-power achieve an error rate of about 40% in a few hundered times realtime on high-end work-stations.

The speech workshop at Johns Hopkins University tries each year to concentrate combined effort of a few researchers from all over the world to improve basic speech recognition technology. The impression of the author was, that this workshop gives every attendee an excellent chance of understanding more about what makes a state-of-the-art speech recognition system, and it gives a great chance of meeting many excellent rearchers working in speech recognition and related areas.

References

[1] L. Deng,, "A dynamic, feature-based approach to speech modeling and recognition", *Proceedings of the 1997 IEEE Workshop on automatic speech recognition and understanding*, pp.107-114, Santa Barbara, CA, 1997.

[2] J. Picone, S. Pike, R. Reagan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards & M. Schuster, "Initial evaluation of hidden dynamic models on conversational speech ", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, TX, submitted, 1999.

[3] H.B. Richards and J. Bridle, "The HDM: A segmental hidden dynamic model of coarticulation", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, TX, submitted, 1999.