SVOX Participation in Blizzard 2007

Johan Wouters

SVOX AG, Switzerland
wouters@svox.com

Abstract

This paper describes the SVOX system architecture and the steps we took to integrate the Blizzard database into our system. The results of the Blizzard evaluation show that SVOX is a leader in small and large footprint unit selection. Analysis of the mean opinion scores for specific sentences shows where our system can improve most. Some recommendations for future Blizzard Challenges are also made.

1. System Description

The SVOX system is based on a flexible and modular architecture. The system contains a multi-purpose text preprocessor and modules for morphological analysis, sentence analysis, word disambiguation, and unit selection. It supports a wide range of markup commands as well as custom lexica and custom text preprocessing rules.

SVOX is a market leader in TTS for navigation systems. One key feature of the SVOX system is its mixed linguality. This feature allows names and places of interest to be spoken in a target language with the correct foreign pronunciation, for example "Beaubourg" is pronounced by the English voice as "bow-bu:r" and not "be:-aw-borg". This is achieved by mapping the native phonetic transcription for all names in the map data of the navigation system to the closest matching phoneme sequence in the target voice.

In the SVOX system, language and voice specific data are fully separated from the runtime code. The language and voice packages are called *lingware modules* (see Fig. 1). The system utilizes one or more lingware modules, allowing fast switching between languages and voices.

SVOX has developed lingware modules for almost 20 languages and 30 voices, and we are constantly adding new modules. Our flexible development environment and system architecture allows for a rapid on-demand creation of customer specific voices and languages. About 50 person hours were invested to integrate the Blizzard database into the SVOX system.

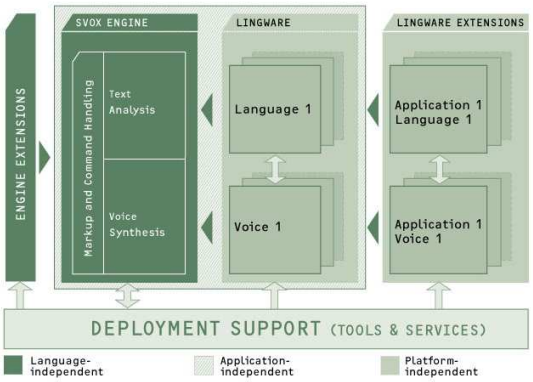

Figure 1: SVOX Architecture.

2. Building Blizzard Voice A

Participants in the Blizzard Challenge 2007 received a database consisting of 6579 utterances sampled at 16 kHz, as well as their corresponding f0 values, phonetic segmentation, and orthographic transcription. Conversion of these data into a SVOX lingware module involved some specific challenges.

We decided not to use the segmentation provided with the database, because it is based on different transcription conventions than are used by the SVOX text analysis module. For example the transcriptions do not mark aspirated plosives. Other mismatching transcription conventions concern glottal stops, r-colored vowels, vowel length, epenthetic stops and approximants, etc. Since the success of a unit selection system depends among others on the match between the transcriptions generated by text analysis and the transcriptions found in the unit database, we did not want to handicap our system by using a database labeled with different transcription conventions.

The transcriptions provided with the Blizzard database also did not come with any guarantees of correctness. In fact, many transcription errors were found. Especially part 3 of the database contains numerous typos, grammatical errors, and Japanese proper names that are not transcribed according to the speaker's pronunciation. The large number of errors in the provided transcriptions was another reason why we did not attempt to normalize them in accordance to our transcription conventions.

2.1. Phonetic transcription

Phonetic transcriptions were generated with the SVOX US English text analysis module based on the orthographic text provided with the Blizzard database. Ideally, a manual revision of all phonetic transcriptions is needed to correct prediction errors and errors due to mismatches between the recorded speech and the orthographic text. Given the limited time available to integrate the Blizzard voice in our system, we decided only to do a manual screening of out-of-vocabulary (OOV) words.

To perform these selective manual transcriptions, we modified an in-house tool that displays the orthographic text of each sentence in a top window, and the phonetic transcription in a bottom window. A play button allows the user to listen to the corresponding recorded sentence or a TTS reading. The tool was extended for the Blizzard Project to highlight OOV words in the orthography window. A button to jump to the next sentence with an OOV word was added.

Using our latest version of the SVOX US text analysis lingware, we found several hundred OOV words in the Blizzard database. A trained linguist required about a day with the extended in-house tool to transcribe all OOV words. In the course of this work, also the orthography was occasionally corrected, i.e. in cases where words were missing, in the wrong order, or mistyped. These orthographic problems were often corrected by the voice talent during the recording, but had remained in the text.

2.2. Phonetic segmentation

The phonetic segmentation is entirely HMM based. Recently, we made significant progress in our automatic alignment by investigating alternative acoustic features, and by fine-tuning design choices such as the number of states and mixtures. In order to combine efficiency with flexibility, we integrated highly optimized C-code for the core HMM algorithms with Python scripting. Together with new features for distributed computing, this resulted in a significant increase in speed and accuracy of our alignment procedure. The training and alignment for the Blizzard database took about 2 hours.

To evaluate the quality of the automatic segmentation, we performed a manual segmentation of the 30 first sentences of the Blizzard database. The automatic boundaries have a 12.9 ms absolute mean error compared to the manual boundaries, with a 95% confidence interval of the mean of 11.8 ms to 14.6 ms. 92% of the automatic boundaries are within a 30 ms range compared to the manual boundaries. No automated post-processing of the HMM labels or further manual segmentation corrections were undertaken.

2.3. Extraction of prosodic features

Besides the standard features such as phonetic identity and context, the SVOX system uses a well balanced mixture of low level and high level prosodic features for unit selection. High-level prosodic features such as phrase breaks are based on a syntactic analysis of the input text.

Low-level prosodic feature extraction is the most complex and time consuming step in our voice building process. We use internally developed algorithms to perform high accuracy pitch and voicing detection and to extract other acoustic features.

2.4. Quality labels

One aspect of the SVOX unit selection approach is to detect off-line units which should be avoided during unit selection. A small percentage of the database is labeled with a negative quality label. The corresponding segments are referred to as *suspect* units. The quality label is based on probability information from the automatic segmentation and statistical outliers with respect to duration, pitch, and other acoustic features. During unit selection, the suspect units are not entirely removed, but they receive a high target cost. This distinguishes the quality labels from other pruning techniques and makes the approach more robust against tagging mistakes. Rare unit sequences are still in the database and can be selected even when they have a negative quality label.

A similar approach to identify suspect units was decribed by Kominek *et al.* at Eurospeech 2005 [1]. There, the outlier information was used to prompt a manual review of suspect units during database preparation. We take the approach a step further by making the outlier information available during unit selection.

In addition to the statistically determined quality labels, the SVOX system also supports manual quality labels. These are used for units that subjectively sound bad and should be avoided during unit selection. The quality labels can be used to manually fine-tune an utterance, by tagging units that have a negative impact on the synthesis quality and thus forcing selection of a different unit. However, the impact of a chosen unit on speech quality is context dependent. Within the scope of a single sentence, one may wish to keep a unit in one context, and avoid it in another. This makes the quality feature less practical for manual fine-tuning, as it stays constant during unit selection.

It is clear that no manual tuning in this or another form was used for the sentences we submitted for the Blizzard evaluation, as this is against the objective of the Challenge.

3. Building Blizzard Voice B

The participants to Blizzard 2007 were invited to submit test sentences for 3 voices. Voice A is the largest voice, and can utilize all material in the Blizzard database.

Voice B consists of the first 1032 sentences of the Blizzard database, out of a total of 6578 sentences. The sentences for Voice B correspond to the Arctic subset, designed by Kominek *et al.* [2]. The total duration of the Arctic sentences is 2914 seconds (48:34 minutes), whereas the total duration of the full data set is about 616 minutes, or 12.7 times more. About 25% of these 616 minutes are labeled as silence.

The Arctic subset was designed to cover an English diphone set by greedy text selection. However, the Arctic corpus is rather small for unit selection, as has been reported in previous editions of the Blizzard Challenge [3]. Although prosodic features are key in unit selection, the design of the segment database was not based on prosodic criteria (see section 4).

To build Voice B, we simply truncated our database for Voice A. We did not rerun the phonetic segmentation using just the sentences in the Arctic subset, as was done by Eide *et al.* [4] for Blizzard 2006. This might have been more to the letter of the Blizzard rules, which state that for Voices B and C no extra material from the database can be used e.g. for prosody. However we decided to keep the segmentation the same for all three voices. The original CMU segmentation accompanying the Blizzard database also does not contain separate labels trained only on the Arctic sentences.

4. Building Blizzard Voice C

The challenge for Voice C was to select a subset of sentences from the full Blizzard data set, based only on the text. No acoustic information from the recordings could be used, except for the duration of the recorded sentences. The sentence durations were used to limit the number of sentences to be selected. Specifically, sentences adding up to maximum 2914 seconds of speech could be selected, which corresponds to the total duration of the Arctic subset as spoken by the database speaker.

We approached this challenge by using the TTS system to transcribe all the sentences phonetically and to predict prosodic patterns. From this we derived prosodically annotated phonetic units, similar to the input features used in unit selection.

The prosodically annotated phonetic units were divided into clusters, resulting in a set of coarse target prosodic-phonetic units. One example of such coarse target prosodic-phonetic units are syllables labeled with a binary prosodic feature for sentence finality.

Then a greedy algorithm was used to select sentences that maximally cover the space of the coarse target prosodic-phonetic units. The greedy algorithm was stopped when the next best sentence exceeded the maximum total duration. One last sentence was selected to fill up the duration budget.

5. Discussion

5.1. Performance in Blizzard evaluation

SVOX is one of the four systems attaining a median MOS score of 4 for Voice A. Natural speech attained a median score of 5. The remaining 12 systems attained a median score of 3 or less for Voice A.

SVOX has the highest MOS score for Voice C. For Voice B, we obtained a shared first place. The differences among the highest scoring systems are not statistically significant. However the results prove our competence especially for small footprint systems.

5.2. Analysis of specific sentences

The participants were requested to submit 100 conversational sentences and 100 news sentences as well as 100 sentences from novels, 50 modified rhyme test and 50 semantically unpredictable sentences. Only 17 conversational sentences and 17 news sentences were used for the actual MOS evaluation of Voices A and B, corresponding to the number of participating systems plus the original speaker. Only 12 conversational and 12 news sentences, a subset of the 17 sentences selected before, were used in the MOS test for Voice C.

We tried to relate the MOS scores we received for these two times 17 sentences to specific strong and weak points of the synthesized waveforms. Although it is difficult to generalize, it seems that prosodic errors such as unnatural rhythm or inappropriate pitch were rated more harshly than segmental errors.

This observation is further supported by the fact that we obtained the same average MOS scores for Voices A and C. When we compare sentences from A and C back to back, Voice A clearly has less concatenation artifacts,

but this did not lead to a difference in the average MOS scores.

We hypothesize that the response of listeners is often based on something that "sticks out" in an utterance, like a wrong phrase break or an unlucky intonation contour. These high level characteristics depend to a large extent on the prosodic description of the sentence going into unit selection, which is shared between Voices A and C.

A difference in smoothness may only affect the MOS score if it is the most important source of unnaturalness, but not when it is dominated by prosodic errors. In a comparative MOS test (CMOS) [5] listeners might be able to focus on differences that are spread out across the utterance, while in a regular MOS test, listeners are biased towards specific problems.

5.3. Dependence on database integrity

Unit selection is a process to match a target linguistic description with unit descriptors found in the database. The synthesis quality depends therefore on the correctness of the unit descriptors. Segmentation errors typically result in disturbing glitches. Transcription errors result in pronunciation problems, i.e. an extra function word or syllable is inserted, vowel quality is wrong, or phonemes are missing.

The unit selection system can be designed to cope with database errors. The quality label described in section 2.4 is an example of this. The use of spectral distances in the selection cost is another example. But even if mislabeled units are avoided during unit selection, they still represent a missed opportunity since useful speech material is lost. Hence labeling errors always decrease the maximum attainable quality for a given unit database.

As a first time participant in the Blizzard Challenge, we were surprised at the amount of transcription errors and the number of foreign words in the provided database. When left uncorrected, these transcription errors are a liability for any unit selection system. We therefore invested in a manual transcription of at least the OOV words. No manual segmentation corrections were undertaken. For future Blizzard Challenges, it would be worthwhile to make a more accurate orthographic and even phonetic representation of the data available to the participants. This could be achieved by sharing corrections among the participants, or by collecting corrections from previous participants if the database remains the same. Alternatively, if the same database is used for subsequent Challenges, participants will have time to improve their transcriptions and segmentation with manual work, thus making it harder for newcomers to enter the Challenge.

6. Conclusion

We decided to participate in the Blizzard Challenge 2007 to test our leadership in large and small footprint unit selection systems, and to find out which aspects of our system could be further improved. We were able to use the results of the MOS test to analyze which sentences scored higher and which were found more problematic. This analysis shows that prosody is more important than segmental accuracy in the MOS results. We look forward to listening to the other high-ranking systems, to hear which sources of unnaturalness they share with our system and which they solve.

7. References

[1] J. Kominek, C. Bennett, A.Black, "Evaluating an correcting phoneme segmentation for unit selection synthesis", Eurospeech 2003, Geneva

[2] J. Kominek, A. Black, "The CMU Arctic speech databases", 5th ISCA Speech Synthesis Workshop, 2004, Pittsburgh

[3] R. Clark, K. Richmond, S. King, "Multisyn voices from Arctic data for the Blizzard challenge", Interspeech 2005, Lisbon

[4] E. Eide et al., "The IBM submission to the 2006 Blizzard text-to-speech challenge", Blizzard Workshop 2006, Pittsburgh

[5] ITU-P.l800, "Methods for subjective determination of transmission quality", Recommendation P.800 International Telecommunication Union (ITU), 1996