ARCHIVES OF ACOUSTICS
32, 1, 67–74 (2007)

BASIC PARAMETERS IN SPEECH PROCESSING

The need for evaluation

Harald HÖGE

Siemens AG, Corporate Technology
Otto Hahn Ring 6, 81739 München, Germany
e-mail: harald.hoege@siemens.com

(*received October 16, 2006; accepted December 19, 2006*)

As basic parameters in speech processing we regard pitch, duration, intensity, voice quality, signal to noise ratio, voice activity detection and strength of Lombard effect. Taking in account also adverse conditions the performance of many published algorithms to extract those parameters from the speech signal automatically is not known. A framework based on competitive evaluation is proposed to push algorithmic research and to make progress comparable.

Keywords: prosodic parameters, VAD, strength of Lombard effect, evaluation.

1. Introduction

In the area of speech recognition, speech synthesis and speaker characterization basic parameters are needed which are crucial for good performance of the systems. In this paper we regard two sets parameters. The first is related to prosody and the second characterizes the acoustic properties of the environment including the impact on the speaker's voice. Pitch, duration, intensity and voice quality are well known "prosodic parameters" representing the first set. As second set we select "environmental parameters": signal to noise ratio (SNR), voice activity detection (VAD) and strength of Lombard effect (SLE).

Traditionally, prosodic parameters are used in speech synthesis. For concatenated speech synthesis [1, 2] a suited voice has to be recorded and annotated with respect to pitch and segment boundaries. Intensity and voice quality often are not regarded. As the manual work for annotation is expensive and time-consuming semi automatic methods for annotation are implemented to decrease the efforts. Nowadays prosodic parameters are used in speech recognition and especially in the field of speaker characterization (see [3–5]). Those systems have to work in general under adverse conditions, what leads to the demand of noise robustness for the algorithms estimating the prosodic parameters. Yet it is well known, that most of these parameters are hard to extract from the speech signal, especially under adverse conditions.

Environmental parameters are used in speech recognition and speaker recognition for noise reduction algorithms. Many approaches exist to estimate SNR and VAD from noisy signals. The SLE parameter is new, but it is known, that the performance of speech recognition systems decreases dramatically for speech with Lombard effect. Here basic research is needed to fully understand the mechanism leading to the Lombard effect and to find a measure for its strength.

Many algorithms have been developed to estimate those basic parameters from the speech signal, but their performance has not been explored sufficiently. There exist no adequate benchmarks which would allow to compare published algorithms and to select the most suited algorithms for extracting the prosodic and environmental parameters for a given task.

Benchmarks have been successfully implemented for whole speech processing systems as recognition or synthesis systems. The pioneering work of benchmarks was performed by DARPA [6] in the field of speech recognition. These activities demonstrated that a suited framework for benchmarking can push substantially progress in technology. Later this "benchmarking technology" was called evaluation and was applied on other speech processing systems as for speaker recognition [7] and for speech synthesis [8]. One of the last successful evaluation campaigns in speech recognition was focused on noise robustness, where the environmental parameters played an important role [9].

Basic elements of the framework of evaluations for speech processing systems are:

- specification of the functionality of a system,
- evaluation criteria, which describe the performance of the system,
- evaluation databases, on which the performance can be tested,
- an organizational framework to perform an evaluation campaign.

For evaluation of the algorithms extracting prosodic and environmental parameters from speech a similar framework is needed. Within the consortium ECESS[(1)] activities to evaluate modules of speech synthesis systems and tools related to speech synthesis have been started. The tools are much in the spirit of tools extracting prosodic and environmental parameters. The first parameter which has been chosen for evaluation was the pitch.

2. The ECESS pitch evaluation framework

The main elements of an evaluation framework for algorithms are similar to those mentioned above for speech processing systems:

- specification of the functionality of the algorithm,
- evaluation criteria, which describe the performance of the algorithm,
- evaluation databases, on which the performance can be tested,
- an organizational framework to perform an evaluation campaign.

[(1)] European Centre of Excellence for Speech Synthesis; www.ecess.eu

Within the first ECESS "tool evaluation campaign" algorithms for pitch detection (PDA) and algorithms for pitch marking (PMA) were evaluated. When starting the evaluation process it turned out that the determination of the functionality of the PDA algorithms was clearly defined by determination the number of pitches per time unit. Yet great discussions started to define the correct position of the pitch mark – the epoch. The specification as described in [8] was finally accepted for setting the reference pitch-marks.

2.1. The evaluation criteria

The evaluation criteria chosen for PDA were the already established criteria (see [10])

- *Gross error high (GEH) and gross error low (GEL)*
  The gross error high (GEH) presents the percentage of voiced speech segments for which the detected pitch is more than 20% higher than the reference pitch (*Estimated_Pitch* > $1.2 * $ *Reference_Pitch*). The gross error low (GEL) presents the percentage of voiced speech segments for which the detected pitch is more than 20% lower than the reference pitch (*Estimated_Pitch* < $0.8 * $ *Reference_Pitch*).
- *Voiced error (VE) and unvoiced error (UE)*
  The voiced error (VE) presents the percentage of voiced speech segments which are misclassified as unvoiced. The unvoiced error (UE) presents the percentage of unvoiced speech segments which are misclassified as voiced were used.

The evaluation criteria used for PMA were defined by the success rate *SR* and the accuracy. *SR* is defined according to [11]:

$$SR = \frac{\left|\left\{x \middle| (x \in \textit{Test}) \wedge (x \in \textit{Ref})\right\}\right|}{|\textit{Ref}|} \cdot 100\%,$$

where *Ref* represents the set of all reference pitch-marks, and *Test* represents the test-set pitch-marks. The number of correct pitch-marks is determined by all test set pitch-marks which are in the tolerance interval of the reference pitch-marks. Replicated pitch-marks, found inside the tolerance interval are not considered as correct pitch-marks. In counting the correct epochs, a maximal tolerance deviation of 20% of the period time T at maximal presumable pitch frequency F ($T = 1/F$) is allowed. The success rate *SR* does not include the errors of missing epochs (deletion errors). To include the deletion error the evaluation criteria "accuracy" was introduced [8].

2.2. Evaluation databases

Two evaluation databases have been designed: a "noisy" database and a "high quality" database (see [11]). The noisy database is publicly available [13]. Unfortunately the high quality database was not yet ready at the time of the evaluation campaign but will be used in later campaigns and will be made available via ELRA/ELDA.

The noisy database is dedicated to evaluate PMA and PDA under adverse conditions. This database consists of parts of the SPEECON speech database [12] which comprises several environmental conditions as the car interior, the office, and living rooms. From this database the recordings of 60 speakers was selected (30 male and 30 female speakers, 16 kHz sampling rate). The database is recorded simultaneously with 4 channels where the first channel (C0) is a close talk microphone and the other microphones (C1, C2, C3) are mounted in a distant between 0.3 m and 4 m. Manual reference pitch marking was performed on the low noise channel C0. The pitch marks were automatically transferred to the other 3 channels.

The high quality database is dedicated to evaluate PDA and PMA algorithms to annotate automatically databases used to generate high quality voices for speech synthesis. Those databases are specified by the deliverable D8 of the TC-STAR project[(2)] . According to these specifications (96 kHz sampling rate, 24 bit accuracy, no reverberations, EEG, close-talk and large membrane microphone) 22 male (11 native UK-English, 11 native German) and 22 female (11 native UK-English, 11 native German) speakers were recorded. A semi-automatic glottal epoch-detection procedure was used to produce and manually correct reference positions of all glottal epochs of four sentences for all recorded speakers.

2.3. The ECESS evaluation campaign

The evaluation campaign was conducted by the University of Maribor (UMB) under the guidance of Zdravko Kacic and Bojan Kotnik. Scripts were provided, which allowed each participating institution to run the evaluation with the noisy database by themselves and to report the results to UMB. At the 7th ECESS workshop held on 5th July 2006 in Maribor the results of the evaluation campaign were reported. Some results are plotted in the curves above. On the PDA evaluation three institutions participated, where the institution “UBC” achieved the best results with respect to the evaluation criterion GEH+GEL. Three institutions participated on the PMA evaluation. Additionally two open source PMAs were evaluated. The results below show clearly that noise and reverberation – present in channel C1, C2, C3 is still a great challenge for accurate pitch marking. During the 7th ECESS meeting discussions about the evaluation criteria started. In the next campaign some modifications on the criteria will be done. This concerns the interrelation between voiced/unvoiced detection and the error rates GEH and GEL. Further the voiced unvoiced detection should be extended to a voiced/unvoiced/non speech detection. This approach leads to an evaluation of an extended VAD detector.

[(2)] See link “documents” of the ECESS web-page.

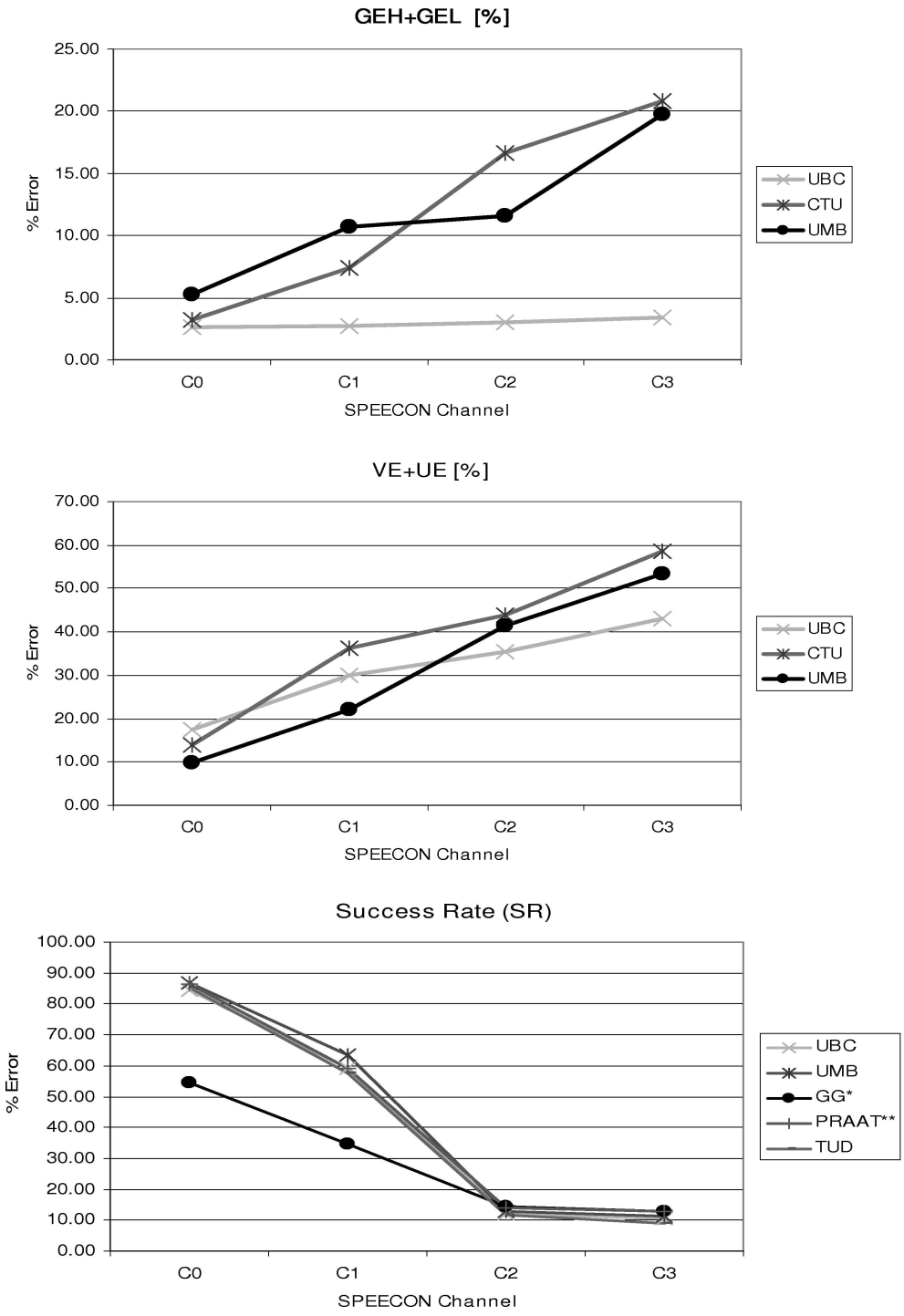

Fig. 1. Error rates of pitch detection (PDA) and pitch marking (PMA) algorithms.

3. Evaluation of other parameters

3.1. Prosodic parameters

Another important parameter is the duration of phonetic units. For concatenative speech synthesis the speech database of a voice has to be segmented into phones. The boundaries of these segments determine the duration of a phone. In phoneme based HMM recognition also the phone boundaries are determined and statistics concerning the duration can be made [5]. The speaking rate is dependent from the duration of syllables and phones within a syllable [14]. Further rhythm is dependent on duration. It is evident, that segmentation and duration are coupled problems. So we have to investigate algorithms, which segment speech into phonetic units automatically. First a reference database is needed which is segmented manually. There exist some databases which are already segmented on phone level (e.g. the TIMIT database), but these databases are not recorded under adverse condition. For segmenting a recorded voice for concatenative speech synthesis mostly the speech is transcribed manually and a voice can be segmented using forced Viterbi alignment. Given such transcribed databases evaluation criteria can be derived. Nevertheless transcription errors have to be taken into account [15]. If the transcription is very error prune or no transcription is available as in speech recognition, segmentation is equivalent to phone recognition. This leads to the task to evaluate a phone-recognizer. In the field of language identification such phone recognizer gains increasing interest [16].

3.2. Environmental parameters

The environmental parameters VAD and SNR have been investigated in many papers relating to noise reduction for speech recognition. VAD is also used in speech coding were standards for public transmission systems exist (e.g. [17]). For evaluation a suited definition of SNR is needed. A specific SNR, which relates SNR-values in certain frequency bands, is proposed in [18], which shows the relation between human and machine speech recognition as a function of SNR. Further suited databases for testing the performance of these parameters have not been set up.

A very specific environmental parameter is the strength of Lombard effect. From related recordings (e.g. [19]) it is evident, that the Lombard effect has acoustic correlates in the speech signal. But to quantify this effect no specification is known. On this topic pioneering work is needed.

4. Conclusion

The paper shows the importance of the parameter selected in speech processing. Approaches to evaluate algorithms have been presented. Still the framework to evaluate all the mentioned parameters, have to be set up.

Acknowledgment

I have to thank Bojan Kotnik who provided the evaluation result of the first ECESS evaluation campaign.

References

[1] HUNT A., BLACK A., *Unit selection in a concatenative speech synthesis system using a large speech database*, Proc. ICASSP 96, 373–376, 1996.

[2] DONOVAN R., ITTYCHERIAH A., FRANZ M., RAMABHADRAN B., EIDE E., VISWANATHAN M., BAKIS R., HAMZA W., PICHENY M., GLEASON P., RUTHERFOORD T., COX P., GREEN D., JANKE E., REVELIN S., WAAST C., ZELLER B., GUENTHER C., KUNZMANN J., *Current status of the IBM trainable speech synthesis system*, Proc. Fourth ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW-4), August 29 – September 1, Perthshire, Scotland 2001.

[3] SHRIBERG E., FERRER L., KAJAREKAR S., VENKATARAMAN A., STOLCKE A., *Modeling prosodic feature sequences for speaker recognition*, Speech Communication, **46**, 3–4, 455–472 (2005).

[4] FERRER L., BRATT H., GADDE V. R., KAJAREKAR S., SHRIBERG E., SONMEZ K., STOLCKE A., VENKATARAMAN A., *Modeling duration patterns for speaker recognition*, Proc. Eurospeech, 2017–2020, 2003.

[5] WILLETT D., GERL F., BRUECKNER R., *Discriminatively trained context-dependent Duration–Bigram models for korean digit recognition*, Proc. Int. Conference on Acoustics, Speech and Signal Processing, ASSP06, pp. I-25–I-28, 2006.

[6] GAROFOLO J. S., FISCUS J. G., FISHER W. M., *Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora*, Proc. DARPA Speech Recognition Workshop, February 1997.

[7] NIST 2005 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf

[8] TC-Star first and second evaluation campaign, www.tc-star.org

[9] HIRSCH H.-G., PEARCE D., *The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions*, Proc. ISCA Tutorial and Research Workshop ASR 2000 – Automatic Speech Recognition: Challenges for the Next Millennium, Paris 2000.

[10] KOTNIK B., HÖGE H., KACIC Z., *Evaluation of pitch detection algorithms in adverse conditions*, Proc. Speech Prosody, 2006, Dresden 2006.

[11] HÖGE H., KOTNIK B., KACIC Z., PFITZINGER H. R., *Evaluation of pitch marking algorithms*, Proc. ITG-Fachtagung Sprachkommunikation. Kiel 2006.

[12] ISKRA D. J., GROSSKOPF B., MARASEK K., VAN DEN HEUVEL H., DIEHL F., KIESSLING A., *SPEECON speech databases for consumer devices: Database specification and validation*, Proc. Second Int. Conference on Language Resources and valuation (LREC'2002), pp. 329–333, Las Palmas 2002.

[13] ELDA catalogue No.: ELDA-S0218; includes construction and short description of the PMA/PDA Reference Database.

[14] PFITZINGER H. R., *Local speech rate perception in German speech*, Proc. of the XIV-th Int. Congress of Phonetic Scienes, Vol. 2, pp. 893–896, San Fransisco 1999.

[15] ADELL J., AGÜERO P. D., BONAFONTE A., *Database pruning for unsupervised building of text-to-speech voices*, Proc. ICASSP, I-889 – I-892, 2006.

[16] MATEJKA P., SCHWARZ P., CERNOCKY J., CHYTIL P., *Phonotactic language identification using high quality phoneme recognition*, Proc. Eurospeech 2005, pp. 2237–2240, Lisbon 2005.

[17] ETSI EN 300 965 V8.0.1 (2000-11) specification: Voice Activity Detector (VAD) for full rate speech traffic channels (GSM 06.32 version 8.0.1 Release 1999).

[18] ANDRASSY B., HÖGE H., *Human and machine recognition as a function of SNR*, Proc. LREC, 2006.

[19] BORIL H., POLLAK P., *Design and collection of Czech lombard speech database*, Proc. ISCA Interspeech, vol. 1, 1577–1580, Lisbon 2005.