suitability of applying regression methods for IQ estimation may be analyzed regarding it as estimating a continuous mathematical function. Finally, Conditioned Random Fields have shown to work well for sequence tagging. As this is related applying an IQ value to each exchange of a sequence of exchanges, those may also increase IQ recognition performance.

References

1. Walker M., Litman D., Kamm C. A., Abella A. PARADISE: a framework for evaluating spoken dialogue agents, in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 1997.

2. Schmitt A., Schatz B., Minker W. MODELING AND PREDICTING QUALITY IN SPOKEN HUMAN-COMPUTER INTERACTION, in *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, USA, 2011.

3. Ultes S., Schmitt A., Minker W. Towards Quality-Adaptive Spoken Dialogue Management, in *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, Montréal, Canada, 2012.

4. Schmitt A., Ultes S., Minker W. A Parameterized and Annotated Corpus of the CMU Let's Go Bus Information System, in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

5. Vapnik V. N. The nature of statistical learning theory, New York, NY, USA: Springer-Verlag New York, Inc., 1995.

6. Ultes S., Schmitt A., Minker W. On Quality Ratings for Spoken Dialogue Systems – Experts vs. Users, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, to appear.

7. Ultes S., ElChabb R., Minker W. Application and Evaluation of a Conditioned Hidden Markov Model for Estimating Interaction Quality of Spoken Dialogue Systems, in *Proceedings of the 4th International Workshop on Spoken Language Dialog System (IWSDS)*, 2012.

УДК 004.93

ON THE CONCEALMENT OF TRANSMISSION ERRORS FOR DISTRIBUTED SPEECH RECOGNITION SYSTEMS

D. Zaykovskiy

Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany. E-mail: dmitry@Zaykovskiy.de

The client-server speech recognition systems face the challenge to provide consistent performance over diverse channel conditions. It is therefore necessary to develop methods which could anticipate the effect of the transmission errors. In this paper we consider an error mitigation approach which does not modify the original data; instead it tries to reconstruct lost information at the receiver via interpolation of successfully transmitted features. Using the packet identification number the DSR server is able to decide unambiguously which packets were lost and which were closest packets received without error. With correctly received packets before and after the burst, error mitigation module can interpolate missing features.

Keywords: distributed speech recognition, transmission error mitigation, interpolation.

О СОГЛАСОВАНИИ ОШИБОК ПЕРЕДАЧИ В РАСПРЕДЕЛЕННЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Д. Зайковский

Университет города Ульма
Германия, 89081, Ульм, Аллея Альберта Эйнштейна 43. E-mail: dmitry@Zaykovskiy.de

В клиент-серверных системах распознавания речи стоит задача обеспечить последовательную работу в различных условиях канала передачи данных. Таким образом, необходимо разработать методы, которые позволят снизить эффект ошибок при передаче данных. В данной работе рассматривается подход для сглаживания ошибки, который не изменяет исходные данные, вместо этого он старается воссоздать потерянную информацию в приемнике с помощью интерполяции успешно переданных функций. Используя номер пакета идентификации сервера DSR, можно однозначно решить, какие пакеты были потеряны, и какие ближе к полу-

чению без ошибок. С помощью правильно принятых пакетов до и после работы модуля сглаживания можно интерполировать недостающие функции.

Ключевые слова: распределенное распознавание речи, сглаживание ошибки передачи, интерполяция.

The days are numbered where we used our mobile phones exclusively for telephone conversation. Today we have access to thousands of different applications and services for our mobile companions and their number is rapidly growing. However, the usability of such services is still hindered by the limited user interface of mobile devices. Speech based user interface could augment standard interface improving quality of the service. The main problem, however, is that reliable large vocabulary speech recognition cannot be done using limited resources of the mobile phones.

The most vividly discussed proposal to overcome this challenge is the principle of Distributed Speech Recognition (DSR). In this approach, the speech recognition process is separated into two parts: a front-end on the client-side and a back-end on the server-side. The front-end extracts characteristic features out of the speech signal, whereas the back-end, making use of the language and acoustic models performs the computationally costly recognition.

Fig. 1 shows system architecture for DSR. The client captures the speech signal using a microphone and extracts features out of the signal. The features are compressed in order to obtain low data rates and transmitted to the server. At the server back-end, the features are decompressed and subjected to the actual recognition process.

To ensure low latency and reduce transmission costs in the context of DSR the usage of a minimal message-oriented UDP protocol is advantageous for the feature transmission [1]. Since UDP does not use acknowledgement technique, it does not generate extra traffic for the retransmission of lost data. However, this means that some sort of error mitigation process has to be carried out on the server side to compensate for transmission losses. The approach which is considered in this paper aims to recover lost data using successfully received information. Using the packet identification number the DSR server is able to decide unambiguously which packets were lost and which were closest packets received without error. With correctly received packets before and after the burst, error mitigation module can interpolate missing features.

Fig. 1 shows system architecture for DSR. The client captures the speech signal using a microphone and extracts features out of the signal. The features are compressed in order to obtain low data rates and transmitted to the server. At the server back-end, the features are decompressed and subjected to the actual recognition process.

To ensure low latency and reduce transmission costs in the context of DSR the usage of a minimal message-oriented UDP protocol is advantageous for the feature transmission [1]. Since UDP does not use acknowledgement technique, it does not generate extra traffic for the retransmission of lost data. However, this means that some sort of error mitigation process has to be carried out on the server side to compensate for transmission losses. The approach which is considered in this paper aims to recover lost data using successfully received information. Using the packet identification number the DSR server is able to decide unambiguously which packets were lost and which were closest packets received without error. With correctly received packets before and after the burst, error mitigation module can interpolate missing features.

2. Classical interpolation methods. In the context of the DSR one can find several interpolation methods for loss recovery. Most prominent among those are the nearest neighbor repetition [2], the linear interpolation and the cubic Hermite polynomial interpolation [3]. In the following we will use notations introduced by James and Milner in the above mentioned work. For the loss burst of the length B with $\mathbf{X}_{\mathbf{before}}$ and $\mathbf{X}_{\mathbf{after}}$ being feature vectors correctly received immediately before and after erasure. The standard feature vector $\mathbf{X}_t$ contains 14 feature components characterizing speech frame *t* [2]. The missing feature vectors $\mathbf{X}_n$ for$1 \leq n \leq B$ will be determined as follows:

– **nearest neighbor repetition** – the missing frame is replaced by the nearest correctly received frame

$$X_n = \begin{cases} X_{\text{before}}, & n < B/2, \\ X_{\text{after}}, & n \geq B/2. \end{cases}$$

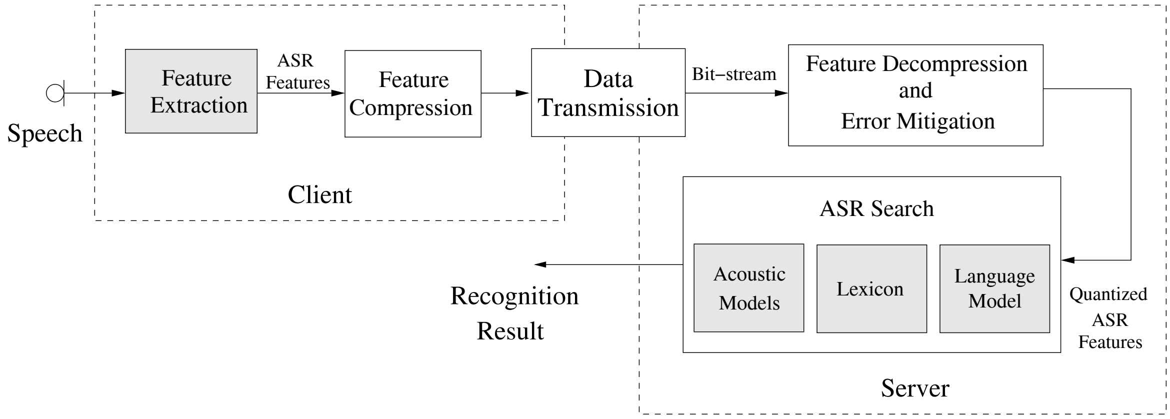


Fig. 1. Client-Server based ASR system – Distributed Speech Recognition

– **linear interpolation** – missing features are interpolated using linear function of time

$$X_n - X_{\text{before}} + \frac{n}{B+1}\left(X_{\text{after}} - X_{\text{before}}\right),$$
$$1 \le n \le B;$$

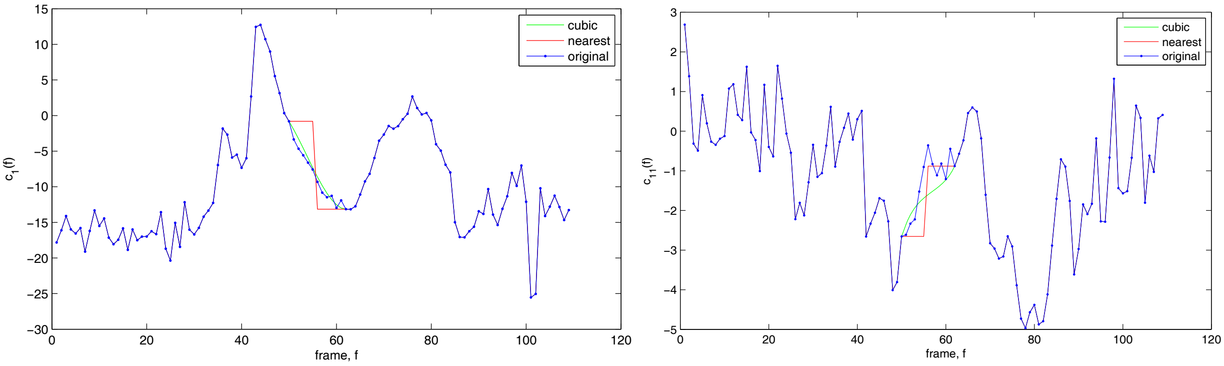
– **cubic Hermite polynomial** – is cubic polynomial interpolator with parameters implying continuous first derivatives of polynomial at the burst edges

$$X_n = X_{\text{before}}\left(1-3t^2+2t^3\right)+$$
$$+X_{\text{after}}\left(3t^2-2t^3\right)+$$
$$+X'_{\text{before}}\left(t-2t^2+t^3\right)+X'_{\text{after}}\left(t^3-t^2\right),$$

where $t = n/(B + 1)$ with $1 \le n \le$ B and X'_{before} and X'_{after} are approximations of derivatives which are iteratively calculated to preserve "nice looking" shape of the interpolated data (we have used Matlab implementation). At this point we have to note that due to the derivatives estimation the cubic interpolator requires two consecutive feature vectors before and after the error burst.

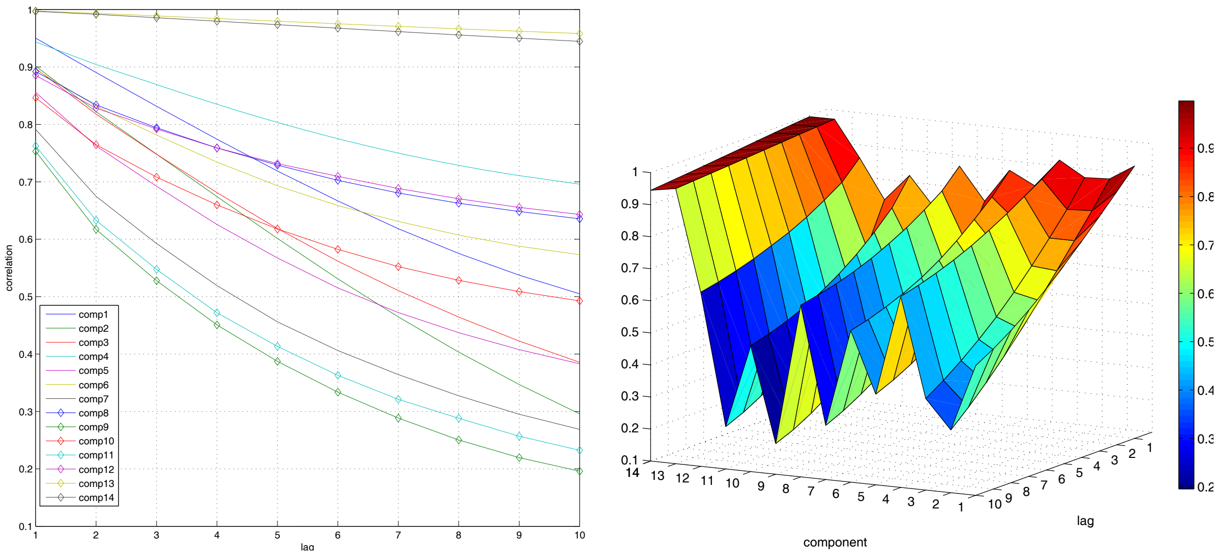
Hybrid Correlation-Based Interpolator. In our experiments we were able to confirm the published results [3], claiming that the advance cubic interpolation to some extend outperforms the simple nearest neighbor repetition. However, a closer analysis of the statistical properties of the cepstral coefficients suggests that not all components can be equally well reconstructed using smooth interpolators, c. f. fig. 2.

In fact when analyzing the interframe component correlation, c. f. Figure 3, one can see that there exist a group of low correlated components and a group of high correlated coefficients. The weak correlated components have very low prediction power. Using sophisticated interpolation therefore does not make any sense for them. In turn, highly correlated coefficients result in a smooth time trajectory which can be well interpolated.



a *b*

Fig. 2. Different interpolators used to recover missing feature components with different correlation level: *a* – 1st feature vector component; *b* – 11th feature vector component



a *b*

Fig. 3. Temporary correlation levels of different feature components: *a* – 2D view; *b* – 3D view

Correlations and threshold x in component classification for hybrid interpolator

	Correlation > x														
Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14	"cubic" component
Intrafame correlation with lag = 3	0,8	0,75	0,75	0,87	0,69	0,78	0,59	0,79	0,53	0,71	0,55	0,79	0,99	0,99	
x = 0.5	x	x	x	x	x	x	x	x	x	x	X	x	x	x	14
x = 0.6	x	x	x	x	x	x	–	x	–	x	–	x	x	x	11
x = 0.7	x	x	x	x	–	x	–	x	–	x	–	x	x	x	10
x = 0.8	x	–	–	x	–	–	–	–	–	–	–	–	x	x	4
x = 0.9	–	–	–	–	–	–	–	–	–	–	–	–	x	x	2
x = 1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0

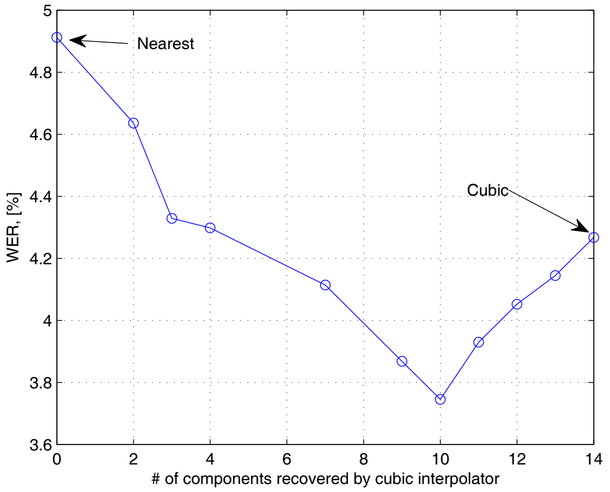
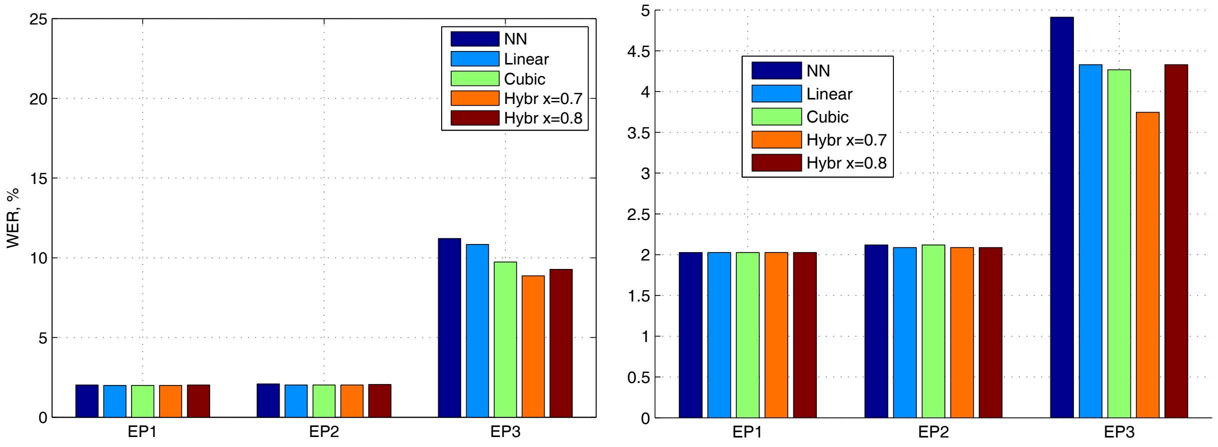


Fig. 4. Dependence of the recognition accuracy on the threshold *x*



a *b*

Fig. 5. Performance of different interpolators in conjunction with different interleaving modes:
a – no interleaving; *b* – packet interleaving with depth *d* = 5

This observation served as a motivation for our hybrid interpolator – it is a combination of the nearest neighbor (NN) approach and the cubic estimator:

– use cubic for highly correlated components – Class I;

– use NN for low correlated component – Class II.

In order to separate the 14 feature components into these two categories a certain threshold has to be found. For that we analyzed 1001 record on the test set 1 from the Aurora-2 database. We have considered interframe correlation with lag three for all 14 components. Table shows the obtained correlations and the classification of components into Class I and Class II for different threshold levels.

It should be noted that when threshold x is set at very low level (x = 0,5) all components are assigned to the Class I and hybrid interpolator converges to a pure cubic approach. For x = 1 it becomes a pure NN interpolator. To determine the optimal threshold level we have performed a number recognition experiments with different threshold levels using real life error pattern EP3.

As we can see on figure 4 the dependence of the recognition accuracy on the threshold shows a clear optimum. Furthermore the optimal WER for the hybrid strategy outperforms individual WERs of both NN and cubic interpolations schemes. The optimal threshold for this test was x = 0,7. It implies cubic interpolation for 10 components and nearest neighbor estimations for the remaining 4.

Experimental Results and Conclusion. In order to evaluate the performance of different interpolation approaches we have conducted a number of recognition experiments on the test set A from the AURORA-2 task [4]. Three different error patters (EP1, EP2 and EP3) corresponding to good, medium and poor transmission channel quality were used. figure *5*, a shows obtained word error rates. figure 5, *b* compares performance of different interpolation approaches when DSR setup is augmented by the interleaving of input data [3]. From the diagram it becomes clear that basically all methods perform at the same level over good (EP1) and medium (EP2) quality channels. At the same time under more demanding transmission channel conditions the suggested hybrid approach offers some additional gain in quality of service.

In this paper we suggested new hybrid approach to the interpolation of lost features combining nearest neighbor repetition and cubic interpolation. The experiments with different system settings and channel conditions have shown that such an interpolator is more advantageous compared to the standard ones. Furthermore it was shown that it can be easily combined with packet interleaving technique as a joint measure for the concealment of transmission errors.

References

1. Zaykovskiy D., Schmitt A. and Lutz M. (2007). New Use of Mobile Phones: Towards Multimodal Information Access Systems. 3rd IET International Conference on Intelligent Environments, Ulm, Germany.

2. Speech Processing, Transmission and Quality Aspects (STQ) (2000). Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm. ETSI Standard ES 201 108.

3. James A. B. and Milner B. P. (2004). Interleaving and estimation of lost vectors for robust speech recognition in burst-like packet loss. In Proc. EUSIPCO, p. 1947–1950.

4. Hirsch H.-G. and Pearce D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proc. ISCA ITRW ASR2000, p. 181–188, Paris, France.