SELF-SUPERVISED MULTI-TASK LEARNING FOR ENHANCED PROSODY PREDICTION IN GERMAN ARTICULATORY SPEECH SYNTHESIS

Zihao Huang, Tianyi Zhang, Peter Birkholz

Institute of Acoustics and Speech Communication, Dresden University of Technology
zihao.huang@mailbox.tu-dresden.de

Abstract: This paper presents a systematic comparison of self-supervised pre-training strategies for prosody modelling. We evaluated three pretext tasks within a unified LSTM-based architecture. The pre-trained encoder is integrated into a multi-task prosody model that jointly predicts phoneme duration, fundamental frequency (f_0), and voicing. Objective evaluation showed that all pre-training methods improve prosody prediction compared to a baseline, particularly for pitch. Subjective listening tests, however, revealed no significant differences in perceived naturalness, indicating that objective gains do not always translate into perceptual advantage. These findings demonstrate that self-supervised pre-training enhances prosody prediction, while perceptual benefits depend on specific aspects of prosodic realization.

1 Introduction

Naturalness has long been a primary objective in the field of text-to-speech (TTS). Prosodic features such as phoneme duration, fundamental frequency (f_0), and energy play an important role for naturalness and intelligibility. Traditionally, these components were modelled separately, for example with Klatt-style rules [1] for duration prediction and multi-space distribution HMMs (MSD-HMM) [2] for f_0 prediction. More recently, end-to-end neural networks such as FastPitch [3] have incorporated prosody modelling within the TTS pipeline to enhance perceptual quality. However, this trend has shifted attention away from fundamental analyses of prosody and the interactions among core parameters such as duration and f_0. For parametric speech synthesis like the articulatory speech synthesizer VocalTractLab (VTL) [4, 5], explicit prosody prediction remains an important topic. For example, VTL allows to synthesize speech by aeroacoustic and articulatory simulation, allowing for precise control of prosodic features like speaking rate, pitch level and voice quality.

Self-supervised learning (SSL), which uses auxiliary pretext tasks to learn rich representation from large amounts of data, has demonstrated strong performance across multiple domains, including natural language processing (NLP) and speech technology [6, 7]. Wav2vec2.0 [8], an extension of contrastive predictive coding (CPC), and HuBERT [9], based on masked prediction of clustered acoustic units, have shown significant improvements in automatic speech recognition (ASR) by exploiting contextualized acoustic representations. Regarding TTS, PnG BERT [10], Mixed-Phoneme BERT [11], and Phoneme-Level BERT [12] use different combinations of inputs, such as phonemes, sub-phonemes, and graphemes, and various pretext tasks to improve pre-training ability. However, due to the limitations of the BERT architecture, these approaches are restricted to non-causal pretext tasks, like masked language model (MLM) and phoneme-to-grapheme (P2G). Moreover, existing studies primarily assess their effectiveness through subjective evaluation, without explicitly examining how pre-training impacts prosodic

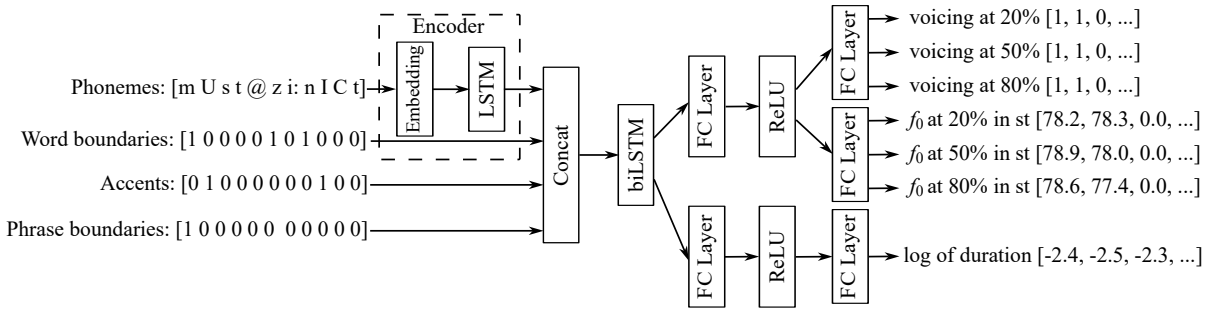

Figure 1 – Outline of the model for prosody prediction with the input data extracted from the German sentence "Musste sie nicht".

predictors. Therefore, it remains unclear which pretext strategy is most beneficial for prosody modelling, particularly for duration and f_0 prediction.

In this work, we propose a multi-task learning (MTL) based prosody model that uses self-supervised pre-training. We make a systematic comparison of contrastive (CPC), generative (LM) and predictive (MLM) SSL strategies under a unified architecture. Experiments show that self-supervised pre-training consistently improves prosody prediction compared to a randomly initialized baseline. Subjective evaluation, however, shows a more nuanced picture, with listener preferences varying across sentences and not always reflecting objective gains.

2 Proposed Method

The proposed framework integrates a pre-trained encoder with a downstream multi-task prosody model. In the first stage, the encoder is trained separately using different pretext tasks to obtain rich, contextualized phoneme representations. In the second stage, the pre-trained encoder is fine-tuned within a prosody prediction model that jointly predicts prosodic parameters through task-specific heads. An outline of the proposed model is illustrated in Figure 1.

2.1 Pre-Training Strategies

The purpose of pre-training is to build a general encoder from large amounts of unlabelled data and to learn contextualized information that can benefit specific downstream tasks. To this end, we investigated three different pretext tasks: Contrastive Predictive Coding (CPC), autoregressive Language Modelling (LM), and Masked Language Modelling (MLM). All methods operated at the phoneme level, where phoneme sequences were used as inputs to a Long-Short-Term Memory (LSTM) based encoder, followed by a task-specific fully connected (FC) layer.

Contrastive Predictive Coding (CPC) learns representations by autoregressively predicting future latent features and contrasting the predicted vector with the true future and sampled negative embeddings. To capture long-range dependencies, the model predicts multiple future steps, with each step using a distinct FC layer. We limited the prediction to three future steps for simplicity. In our adaptation for phoneme-level modelling, the limited phoneme vocabulary makes standard negative sampling and Noise Contrastive Estimation Loss (InfoNCE Loss) impractical. Instead, we measured the cosine similarity between predicted and target embeddings and optimized a cross-entropy loss over all possible phonemes. As shown in SimCLR [13], the large number of negatives helps in learning more discriminative and robust features. The loss function of modified CPC for each timestamp t is defined as follows:

$$\mathscr{L}_{CPC} = -\sum_{k=1}^{K} \log \frac{\exp(\mathrm{sim}(W_k c_t, e_{t+k})/\tau)}{\sum_{i=1}^{V} \exp(\mathrm{sim}(W_k c_t, e_i)/\tau)}, \tag{1}$$

where c_t is the LSTM hidden state at the time t, W_k the k-th step predictor to be trained, V

the phoneme vocabulary, e_{t+k} the phoneme embedding of k-th future step, sim$(\cdot)$ the cosine similarity, and τ the temperature to control the smoothness of the probability distribution. The phoneme embeddings e_i are jointly learned during training.

The **Language Modelling (LM)** pre-training strategy predicts the next phoneme token in a sequence given the preceding phonemes, operating in an autoregressive manner similar to CPC. Unlike CPC, LM directly maximizes the likelihood of the next token over the entire vocabulary without any embedding normalization or temperature scaling and predicts only one future step. The LM loss function is the conventional cross-entropy loss between the predicted token probabilities and the true phoneme labels described below:

$$\mathscr{L}_{LM} = -\log P(x_{t+1} \mid x_{<t}) = -\log \frac{\exp(W c_t)}{\sum_{i=1}^{V} \exp(e_i)}. \tag{2}$$

Here, W is the weights matrix to be trained. The loss $\mathscr{L}_{LM}$ corresponds to the negative log-likelihood of predicting the next phoneme $x_{t+1} = W c_t$ given the preceding context $x_{<t}$.

Masked Language Modelling (MLM) predicts masked phoneme tokens using the context from both past and future phonemes. Unlike CPC and LM, MLM is a non-causal and predictive method. In our implementation, 15 % of phonemes in the dataset are randomly masked. MLM leverages bidirectional context and focuses on direct reconstruction. The loss function for MLM is a cross-entropy loss only over the masked phoneme positions as shown below:

$$\mathscr{L}_{MLM} = -\log P(x_m \mid x_{\setminus M}) = -\log \frac{\exp(W c_m)}{\sum_{i=1}^{V} \exp(e_i)}, \tag{3}$$

where $x_{\setminus M}$ denotes the input excluding the masked tokens.

2.2 Multi-Task Prosody Model

The proposed prosody model jointly predicts duration, f_0, and voicing within a multi-task learning framework. The input is a phoneme sequence, which is encoded using the pre-trained encoder described above. Prosodic features, including the start of word, word accent, and start of phrase, each represented as a binary indicator, are concatenated with the encoder outputs and passed through a bidirectional LSTM (biLSTM) to capture prosodic-related contextual features. Two FC heads are attached to the biLSTM output: one dedicated to duration prediction and the other to pitch. After a ReLU activation, the duration head performs regression through a final FC layer. The pitch head predicts both continuous f_0 values and voicing. One FC layer predicts three f_0 values at 20 %, 50 %, and 80 % of the phoneme duration, while another FC layer outputs three binary voicing labels at the same positions. The voicing labels are derived from acoustic pitch extraction and serve to mask unvoiced regions, ensuring that f_0 regression is trained only where reliable pitch values exist. The overall loss is a weighted combination of mean square errors of duration and voiced pitch (MSE$_d$ and MSE$_p$) and the cross-entropy loss for voicing classification (CE$_v$):

$$\begin{aligned} \mathscr{L} =& w_d \cdot \mathrm{MSE}_d + \\ & (1 - w_d) \cdot (w_p \cdot \mathrm{MSE}_p + (1 - w_p) \cdot \mathrm{CE}_v), \end{aligned} \tag{4}$$

where w_d and w_p denote the loss weights regarding duration and f_0, tuned as hyperparameters to balance the subtasks.

3 Experimental Setup

3.1 Dataset

We used two corpora. The first corpus was the open-source HUI-Audio-Corpus-German [14], which contains paired audio and text data. From the clean, high-quality subset, we selected 55 unique books recorded by 5 main speakers. Long sentences were segmented at pauses longer than 0.5 seconds, resulting in a total of 67 429 sentences. Phoneme sequences were extracted from the paired audio and text files using WebMAUS BASIC [15, 16]. We adopted the narrow SAMPA transcription, which accurately represents the actual pronunciation of sounds, closely tied to the acoustic domain and provides prosodic distinctions that are more suitable for our task.

As the second corpus, we used the BITS Unit Selection Synthesis Corpus [17] consisting of 1683 German sentences recorded by four professional speakers. Following our prior work [18], we used only the recordings of a single male speaker. All recordings were segmented into phonemic units and annotated with prosodic information.

To ensure consistency between the two corpora, we aligned the phoneme inventories to construct a unified dictionary of 66 phonemes. During alignment, affricates were divided into their constituent consonants, diphthongs produced by WebMAUS as vowel-tiefschwa combinations were merged, and all pauses were removed.

Phoneme Pre-training Data: For the pre-training stage, we used both the HUI and BITS Corpus. Only the phoneme sequences were used, without any prosodic target information. This setup allows the encoder to learn contextualized phoneme representations in a purely self-supervised manner.

TTS Fine-tuning Data: In the downstream prosody prediction task, phoneme sequences from the BITS Corpus were used as inputs. The f_0 values were extracted from the corresponding audio using Praat [19] with filtered autocorrelation ranging from 50 to 400 Hz. Each phoneme was assigned f_0 values at 20 %, 50 %, and 80 % of its duration. Rather than strictly zeroing all unvoiced phonemes, the f_0 target was determined by acoustic pitch extraction to preserve voicing information during coarticulatory transitions. Durations were represented in log scale, while f_0 was converted to semitones with a reference of 1 Hz. Finally, both preprocessed duration and f_0 values were normalized using min-max scaling.

3.2 Training Strategy

All experiments were implemented in Python using the Pytorch [20] framework and trained on a single NVIDIA H100-SXM5 Tensor Core GPU in a high-performance computing environment.

In preliminary experiments, we tuned the base model from scratch and selected the best-performing architecture with an embedding dimension of 28, encoder LSTM hidden size of 64 with 2 layers and dropout rate of 0.49, biLSTM hidden size of 64 with 2 layers and dropout rate of 0.185 for the downstream prosody task, FC layer sizes of 16 for duration and 64 for pitch. The hyperparameter optimization was carried out using Optuna [21] with Bayesian search. This architecture was fixed for all subsequent experiments to ensure fair comparison across different pre-training strategies. For MLM, the encoder hidden size was halved to keep the model size consistent with CPC and LM.

For pre-training, the encoder was trained on the entire dataset. The best checkpoint was selected once the model had sufficiently converged. For fine-tuning, the dataset was split into training, validation, and test sets with an 80/10/10 ratio. To mitigate catastrophic forgetting, the encoder was fine-tuned with a learning rate set to half of the baseline value. Model selection

was based on the mean RMSE (root mean square error) of duration and f_0 across the validation set. Standard deviations of duration and f_0, computed from the training set, were used for the mean RMSE calculation.

3.3 Objective Evaluation

For phoneme duration and f_0 regression, we used the RMSE on each domain to select and evaluate the prediction models. The RMSE of phoneme duration was calculated between target duration d_{true} and the predicted duration d_{pred} per phoneme in milliseconds (ms):

$$E_{\text{dur}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(d_{\text{true}}[n] - d_{\text{pred}}[n])^2}, \tag{5}$$

where N is the number of phonemes in the dataset. With respect to f_0, we calculated the RMSE between the target and predicted f_0 in semitone (st) of the voiced segments only:

$$E_{f_0} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(12 \times \log_2\left(\frac{f_{0_{\text{pred}}}[m]}{f_{0_{\text{true}}}[m]}\right)\right)^2}, \tag{6}$$

where m is the index of the f_0 value and $M = 3N$ is the total number of voiced f_0 values in the dataset.

3.4 Perceptual Evaluation

To assess the perceptual impact of different pre-training strategies, we conducted a listening test with native German speakers. A total of 20 sentences were randomly selected from the Kiel Corpus [22], which were never shown in the training and testing of the models. The stimuli were synthesized using an extended version of the articulatory synthesizer VocalTractLab 2.3. Prosody was predicted by the proposed models under a baseline without pre-training (Base) and three different pre-training strategies (CPC, LM and MLM), resulting in 80 synthetic stimuli in total. All audio samples were normalized to a loudness level of -23 LUFS to ensure comparability and are available in the supplemental material [1].

The listening test was implemented in Praat and followed a forced-choice AB test design. Participants were presented with pairs of stimuli and asked to select the version that sounded more natural. Repetitions were allowed during the experiments. Each participant completed 12 unique comparisons per sentence, resulting in 240 trials per listener.

We recruited 17 native German speakers (5 female, 12 male), aged between 24 and 57 years. Participants reported normal hearing and no known speech or language disorders.

4 Results and discussion

Table 1 summarizes both objective and subjective evaluation results. Objectively, all three pre-training strategies outperformed the baseline in terms of mean RMSE of prosody prediction. When analysing duration and pitch separately, different tendencies appear. CPC and LM, as causal prediction methods, reduce the RMSE of f_0 but at the expense of slightly higher duration error. In contrast, MLM, which leverages bidirectional context, achieves consistent improvements in both duration and pitch, yielding the lowest mean RMSE overall and the highest accuracy in voicing classification. For subjective evaluation, LM received the highest listener

[1] https://www.vocaltractlab.de/index.php?page=birkholz-supplements

preference, even though it did not achieve the best objective metrics. However, statistical analysis using the Wilcoxon signed-rank test showed no significant differences between any pre-training strategy and the baseline.

The objective-subjective disconnect suggests that the minor numerical improvements likely fall below the perceptual threshold for listeners under the current experimental setup. The evaluation stimuli may have represented short and prosodically easy utterances, where the baseline already achieves high adequacy, leaving limited room for SSL-enhanced models to show their superior long-range modelling capabilities. In addition, qualitative feedback from participants indicated that distinct prosodic realization of the same sentence often appeared equally natural. This highlighted the one-to-many mapping inherent in prosody, suggesting that distance-based metrics like RMSE do not fully capture the plausibility of valid prosodic variations. Taken together, the results show that SSL-based pre-training enhances objective prediction quality, but its perceptual advantages remain modest and may depend on specific aspects of prosodic realization.

Table 1 – Subjective and objective evaluation of the best performing models.

	Objective evaluation				Subjective evaluation	
Model	RMSE f_0 (st)	RMSE duration (ms)	Voicing mean	Voicing accuracy (%)	Preference frequency	Ranking
Base	2.197	21.37	1.320	92.49	1047	2
CPC	2.174	21.79	1.319	92.45	968	4
LM	2.165	21.49	1.311	92.49	**1067**	**1**
MLM	**2.145**	**21.31**	**1.300**	**92.62**	998	3

5 Conclusion and Future Work

This study compared contrastive, generative, and predictive self-supervised pre-training strategies for prosody prediction within a unified model architecture. Objective results show that pre-training consistently benefits prosody modelling, especially for pitch prediction, with MLM achieving the strongest overall improvements. Subjective evaluation, however, revealed no significant differences in listener preferences between the baseline and any of the pre-training methods, highlighting a gap between objective gains and perceptual outcomes. A limitation of this study is that only 1683 sentences of a single speaker and 20 short stimuli were used for the final training and subjective evaluation, respectively. Future work will explore larger datasets, more diverse evaluation protocols, and alternative model designs to bridge this gap and better capture perceptually relevant aspects of prosody in articulatory synthesis.

6 Acknowledgements

We acknowledge that this project is co-funded by the European Union and co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament and by the "Zentrales Innovationsprogramm Mittelstand (ZIM)" by the German Federal Ministry for Economic Affairs and Energy (BMWK), grant no. KK5049503FG3. We also acknowledge the computing time made available on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Education and Research and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner).

References

[1] KLATT, D. H.: *Linguistic uses of segmental duration in English: Acoustic and perceptual evidence*. *The Journal of the Acoustical Society of America*, 59(5), pp. 1208–1221, 1976.

[2] TOKUDA, K., T. MASUKO, N. MIYAZAKI, and T. KOBAYASHI: *Multi-space probability distribution HMM*. *IEICE TRANSACTIONS on Information and Systems*, 85(3), pp. 455–464, 2002.

[3] ŁAŃCUCKI, A.: *FastPitch: Parallel text-to-speech with pitch prediction*. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6588–6592. IEEE, 2021.

[4] STONE, S. and P. BIRKHOLZ: *Monophthong vocal tract shapes are sufficient for articulatory synthesis of German primary diphthongs*. *Speech Communication*, 157, p. 103041, 2024.

[5] BIRKHOLZ, P.: *Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis*. *PLOS ONE*, 8(4), pp. 1–17, 2013. doi:10.1371/journal.pone.0060603.

[6] GUI, J., T. CHEN, J. ZHANG, Q. CAO, Z. SUN, H. LUO, and D. TAO: *A survey on self-supervised learning: Algorithms, applications, and future trends*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), pp. 9052–9071, 2024.

[7] MOHAMED, A., H.-Y. LEE, L. BORGHOLT, J. D. HAVTORN, J. EDIN, C. IGEL, K. KIRCHHOFF, S.-W. LI, K. LIVESCU, L. MAALØE ET AL.: *Self-supervised speech representation learning: A review*. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), pp. 1179–1210, 2022.

[8] BAEVSKI, A., Y. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A framework for self-supervised learning of speech representations*. *Advances in Neural Information Processing Systems*, 33, pp. 12449–12460, 2020.

[9] HSU, W.-N., B. BOLTE, Y.-H. H. TSAI, K. LAKHOTIA, R. SALAKHUTDINOV, and A. MOHAMED: *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 3451–3460, 2021.

[10] JIA, Y., H. ZEN, J. SHEN, Y. ZHANG, and Y. WU: *PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS*. *arXiv preprint arXiv:2103.15060*, 2021.

[11] ZHANG, G., K. SONG, X. TAN, D. TAN, Y. YAN, Y. LIU, G. WANG, W. ZHOU, T. QIN, T. LEE ET AL.: *Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech*. In *Proc. Interspeech 2022*, pp. 456–460. 2022.

[12] LI, Y. A., C. HAN, X. JIANG, and N. MESGARANI: *Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions*. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

[13] CHEN, T., S. KORNBLITH, M. NOROUZI, and G. HINTON: *A simple framework for contrastive learning of visual representations*. In *International Conference on Machine Learning*, pp. 1597–1607. PmLR, 2020.

[14] PUCHTLER, P., J. WIRTH, and R. PEINL: *Hui-audio-corpus-german: A high quality tts dataset*. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 204–216. Springer, 2021.

[15] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In J. J. OHALA (ed.), *Proceedings of the XIVth International Congress of Phonetic Sciences : ICPhS 99 ; San Francisco, 1 - 7 August 1999*, pp. 607 –610. San Francisco, 1999. URL `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-13682-6`.

[16] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017.

[17] ELLBOGEN, T., F. SCHIEL, and A. STEFFEN: *The BITS speech synthesis corpus for German*. *age*, 47(45), p. 40, 2004.

[18] STEINER, P., Z. HUANG, A.-L. FIETKAU, and P. BIRKHOLZ: *Neural prosody prediction for german articulatory speech synthesis*. In *Proceedings of the 16th ITG Conference on Speech Communication (ITG-Fb. 321)*, pp. 93–97. VDE Verlag, Berlin, Germany, 2025.

[19] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. *Glot. Int.*, 5(9), pp. 341–345, 2001.

[20] PASZKE, A., S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, A. DESMAISON, A. KOPF, E. YANG, Z. DEVITO, M. RAISON, A. TEJANI, S. CHILAMKURTHY, B. STEINER, L. FANG, J. BAI, and S. CHINTALA: *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX, and R. GARNETT (eds.), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

[21] AKIBA, T., S. SANO, T. YANASE, T. OHTA, and M. KOYAMA: *Optuna: A Next-generation Hyperparameter Optimization Framework*. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[22] KOHLER, K. J.: *Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech*. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3, pp. 1938–1941. IEEE, 1996.