INFORMATICA, 2001, Vol. 12, No. 3, 477–486

Synthesized Fricative *ch* Specific Features and Influence on Speech Quality Analysis

Marija ŽINTELYTĖ, Irmantas KANDRATAVIČIUS

Kaunas University of Techonology
Studentų 50, 3028 Kaunas, Lithuania
e-mail: marija.zintelyte@telecom.lt

Received: February 2001

Abstract. One of speech synthesis main problems is synthesis of unvoiced fricatives. One of our previously stated conclusions is that consonant x is influenced by before and behind existing phonetic elements. The aim of experiments described in this paper is to evaluate influence of different x allophones for speech intelligibility and automatic speech recognition.

In this paper the formal system, which describes allophones and, at the same time, phonemes interrelations in their possible sequences in natural language, is described. The formal system is necessary for automatic speech synthesis questions' solution. The experiments of two different types were carried out in order to evaluate the resemblance between two different ω^x allophones: a) ω^x allophones resemblance analysis based on expert evaluation; b) ω^x allophones resemblance analysis based on automatic speech recognition results evaluation.

Experiment's results corroborated that *ch* allophones differ and depend from the context, i.e., from neighboring vowels, different *ch* allophones have influence on speech intelligibility, and therefore different *ch* allophones for high quality speech must be synthesized.

Key words: speech synthesis, unvoiced fricatives, formal system, expert evaluation, speech recognition, context influence, speech quality.

1. Introduction

Term "dialog" till 20^{th} century sixth decade was understood strictly as two people conversation. Information exchange between man and his work instruments was not in progress.

The first phase of man communication with machinery can be considered as use of various measuring elements with which help man was informed about machinery work processes stages. One of the main man-machine communication problems is organization of man and his working instrument communication using human speech. This can be achieved applying automatic speech recognition and synthesis technologies (Žintelytė, Kandratavičius, 1997; Žintelytė, 1999).

Lately after violent burst of computer use in various man activity spheres, effective speech synthesis and recognition technologies demand became very high.

It has to be mentioned, that current speech synthesis systems do not deliver the quality that customer demands. They sound unnatural and their speech is not pleasant to listen.

Therefore one of the main tasks to all languages' speech synthesis systems – synthesized prosody and voice quality improvement (Sonntag *et al.*, 1999).

In order to ensure the required synthesized speech quality, specific features of synthesized language and its phonetic units must be evaluated.

It is evident, that all above mentioned pressing problems actual for Lithuanian speech synthesis too. While solving these problems specific features of Lithuanian language must be taken into account.

Acoustic features of Lithuanian consonants are very little investigated (Pakerys, 1995). In order to carry out high quality Lithuanian speech synthesis, lack of information about consonant acoustic features is confronted.

One of speech synthesis main problems is synthesis of unvoiced fricatives, e.g., ch, h, f.

Fricative *ch* (further x), that is one of the most complicated unvoiced fricatives, was chosen as object for further research and analysis.

Consonant x, like consonants h and f, are new Lithuanian language consonants. They came to Lithuanian language from words with foreign origin. They are not frequently used in Lithuanian speech. In order to perform high quality Lithuanian speech synthesis, characteristics of above-mentioned consonants must be analyzed and evaluated, because synthesis of these consonants in some spheres (e.g., surnames of Lithuania residents, international words) is an obligatory condition.

As we know, there are no formants in spectrograms of unvoiced consonants (Girdenis, 1981). The analysis results display that some higher energy areas, which vary in duration, frequency and energy value, were observed in the spectrograms (Žintelytė, 2001). Continuous frequency area we named as basic frequency. Basic frequency was used for fricatives specific features research.

One of our previously stated conclusions is that consonant x is influenced by before and behind existing phonetic elements. This conclusion was made after consonant x phonetic analysis has been carried out. Phoneme x basic frequency is altering depending on before and behind standing vowel tone pitch. Performing high quality Lithuanian speech synthesis, it is necessary to synthesize different x allophones that depend on the context. Different x allophones depend on the vowel tone pitch (Žintelytė, 2001). In that case different x allophones vary in their basic frequency.

Having in mind, that x allophones depend on their basic frequency, next conclusion can be drawn: allophones acoustically differ the more, the more greater is the difference between comparative x allophones basic frequencies. This difference must have influence for synthesized speech intelligibility.

The aim of experiments described in this paper is to evaluate influence of different x allophones for speech intelligibility and automatic speech recognition.

2. Formal Grammar

For automatic speech synthesis questions' solution it is expedient to have formal system, which describes allophones and, at the same time, phonemes interrelations in their possible sequences in natural language.

Next marking is introduced.

- Primary language alphabet $A_p = \{\alpha^p\}$, where $\alpha_i^p \in A_p$ is primary alphabet letter, i.e., at the same time it is also language alphabet letter.
- It is expedient for Lithuanian language to introduce to the alphabet used in formal system combinations of the primary alphabet letters, i.e., α^p combinations. These combinations are formed referring to research recommendation (Žintelytė, 2000). Set of these combinations is $A_s = \{\alpha^s\}$, where $\alpha_i^s \in A_s$ is primary language alphabet letters combination.
- For further investigations next language alphabet will be used $A = A_p \cup A_s$, where $A = \{\alpha\}$, and $\alpha_i \in A$ can be either language alphabet letter, either combination of these letters. Further α_i in general will be named as alphabet A element or, where this will not cause ambiguity, simply element. Investigated language alphabet (further alphabet) is $A = \{\alpha\}$, where $\alpha_i \in A$ is alphabet element.
- Set of alphabet A phonemes is $A^* = \{\alpha^*\}$, where $\alpha_i^* \in A^*$ is phoneme of element α_i.
- Phoneme α_i^* beginning of basic frequency $\xi^- (\alpha_i^*)$.
- Phoneme α_i^* ending of basic frequency $\xi^+ (\alpha_i^*)$.
- Phoneme α_i^* duration $\tau (\alpha_i^*)$, tone pitch$\rho (\alpha_i^*)$.

All above-mentioned phoneme α_i^* parameters are alternating. Allophones describe their exact values.

Allophones set used in language is described as $\Omega = \{\omega\}$, where $\omega_j = \left\{\omega_j^i\right\} \in \Omega$, here $\omega_j^i = \left(\alpha_i^*; \xi_j^- (\alpha_i^*); \xi_j^+ (\alpha_i^*); \tau_j (\alpha_i^*); \rho_j (\alpha_i^*)\right)$.

So $\|\Omega\| > \|A^*\|$, where $\|X\|$ means power of X set – set elements number.

Thus vector ω_j^i defines concrete allophone, i.e., phoneme expressed by concrete sound, which in writing is represented by element.

While speaking allophone sequences in time are used. If discrete time axis $t = 1, 2, ..., \lambda$ will be used, where λ is finite size and each analyzed time axis point will be identified with allophone pronunciation and will be marked as $\omega(t)$, then we will have allophone sequence during time axis describing speech fragment: $\omega(1)$, $\omega(2)$, ..., $\omega (t-1), \omega (t), \omega (t+1), \ldots$.

It is very important to know allophone dependence upon context for automatic high quality speech synthesis, i.e.,

$$\omega (t) = f (\omega (t-1), \omega (t+1)),$$

where $\omega (t-1)$ and $\omega (t+1)$ joins influences before and after $\omega(t)$ existing allophone sets, i.e., encompasses history and future.

Detailing above presented dependency for allophone ω_j^i we have:

$$\xi_j^- (\alpha_i^*) (t) = \varphi_j (\tau_j (\alpha_i^*) (t-1), \rho_j (\alpha_i^*) (t-1));$$

$$\xi_j^+ (\alpha_i^*) (t) = \varphi_j (\tau_j (\alpha_i^*) (t+1), \rho_j (\alpha_i^*) (t+1)).$$

Presented expressions generalize main allophones characteristics fixed in spectrograms and assists in formal investigation of Lithuanian allophones beginning and ending basic frequency dependency upon alongside standing allophones tone pitch and duration.

Applying above presented grammar and taking into account x allophone beginning and ending basic frequencies (which impact x acoustic features mostly (Žintelytė, 2001)) analyzed x allophones described as:

$$\omega_j^x = \left(\alpha_x^*,\ \xi_j^-\left(\alpha_x^*\right),\ \xi_j^+\left(\alpha_x^*\right)\right).$$

Allophone ω^x will be investigated in the context of Lithuanian vowels. Vowel influence will be analyzed taking into account Lithuanian vowel's tone pitch. Vowel set sorted out according tone pitch: *u, o, a, e, ė, i*. Further we will consider that vowels arranged in this sequence differs one from another in one positional row according tone pitch (further positional row). Dividing line between high and low tone pitch vowels passes between a and e.

Analyzed combinations $\left(\omega^m\left(t-1\right),\omega^x\left(t\right),\omega^n\left(t+1\right)\right)$, where $\omega^m\left(t-1\right)$ and $\omega^n\left(t+1\right)$ – Lithuanian vowel allophones, when $\omega^m\left(t-1\right)$ and $\omega^n\left(t+1\right)$ can also be the same vowel allophones. In that way 36 different combinations and 36 different ω^x are received (36ω^x allophones grammar).

After estimation of precondition, it can be asserted that in these 36 combinations ω^x allophone can be decomposed into two phonetic units, depending on $\omega^m\left(t-1\right)$ and $\omega^n\left(t+1\right)$. These phonetic units can be treated as separate ω^x allophones. Thus 6 different ω^x allophones (6 ω^x allophone grammar) are received, based on which 36 above-mentioned combinations can be formed: $\left(\omega^m\left(t-1\right),\omega^{x'}\left(t\right),\omega^x\left(t\right),\omega^n\left(t+1\right)\right)$, where $\left(\omega^{x'}\left(t\right),\omega^x\left(t\right)\right) = \omega^x\left(t\right)$, when $\omega^{x'}\left(t\right)$ and $\omega^x\left(t\right)$ can also be the same ω^x allophones.

ω^x positional row can be described in the same way as vowel positional row according tone pitch, i.e., $\omega^u\left(t-1\right),\omega_1^x\left(t\right),\omega^u\left(t+1\right);\ldots;\omega^m\left(t-1\right),\omega_k^x\left(t\right),\omega^m\left(t+1\right);\ldots;\omega^i\left(t-1\right),\omega_6^x\left(t\right),\omega^i\left(t+1\right)$, where k – positional row index.

This paper deals with experiments, which investigate 6ω^x and 36ω^x allophones grammars.

3. Experiments

3.1. *Goals*

The main experiment goal – different unvoiced consonant ω^x allophone resemblance evaluation.

The experiments of two different types were carried out in order to evaluate the resemblance between two different ω^x allophones:

1. ω^x allophones resemblance analysis based on expert evaluation.

2. ω^x allophones resemblance analysis based on automatic speech recognition results evaluation.

First type experiments, i.e., experiments carried out with real or synthetic (artificially created neologisms) polysyllabic words, certainly containing three sounds (ω^x in vowels context) combinations. In these words speaker pronounced ω^x allophones resemblance analysis was carried out referring to expert evaluation.

During second type experiments ω^x allophones resemblance evaluation was carried out using automatic speech recognition system, based on hidden Markov models (HMM). Resemblance between different ω^x allophones was estimated analyzing automatic speech recognition system's errors reasons, performing analysis of mostly confused ω^x.

3.2. *Expert Evaluation*

Ten (10) listeners participated in the experiments. Listeners listened to 108 (3 experiments with 36 polysyllabic words each) records. One record is compiled from primary and secondary words pair. Words contain above-mentioned three sounds combinations. Primary word in the experiments is named as word pronunciation standard. Secondary word ω^x (further substitution ω^x) was modified by ω^x, that was transferred from specially selected word (further transferable ω^x), which ω^x context corresponds to specific requirements, e.g., record: *ša*ω_1^x*as*, *ša*ω_2^x*as*; where transferable ω_2^x is taken from word *psi*ω_2^x*ika*. Here substitution ω_1^x is word *šachas* standard consonant.

Listeners listened and compared primary and secondary words, where the latter are primary words after substitution, i.e., primary words with from the other context taken transferable ω^x. Conclusions were stated after generalization of listeners' estimation.

Change of secondary word ω^x pronunciation in comparison with standard ω^x (i.e., primary word) after ω^x alteration was evaluated. Listeners their evaluations presented by filling questionnaire. Listeners' secondary word ω^x correspondence to primary word ω^x estimated in five-level system. Used estimations: excellent (1); good (2); satisfactory (3); bad (4) and very bad (5).

During experiments it was tried to determine how listener reacts to transferable ω^x, taken from possible border variants – the highest and the lowest tone vowel context, and then transferable ω^x is taken from vowel a context, i.e., $(\omega^a(t-1), \omega^x(t), \omega^a(t+1))$ context), which stands in the middle of positional row. Results were generalized calculating estimations' arithmetical means. Part of generalized experiments' results (shown only cases, when before and after substitution ω^x existing vowels are identical) is presented in Table 1 (where, e.g., x is taken from *uxu* means $\omega^x$is taken from $(\omega^u(t-1), \omega^x(t), \omega^u(t+1))$ and x means ω^x).

Next conclusions were drawn after experiments results' generalization.

1. When transferable ω^x is taken from high tone vowel context, listener listening to secondary word hears lower quality pronunciation, than transferable ω^x is taken from low tone vowel context. Listeners estimations:
   - transferable ω^x is taken from context $(\omega^u(t-1), \omega^x(t), \omega^u(t+1))$ – 42% of records estimated <=2;

Table 1
Evaluation results

Evaluation	*uxu*	*oxo*	*axa*	*exe*	*ėxė*	*ixi*
x is taken from *uxu*	1.0	1.3	2.3	2.3	2.7	4.7
x is taken from *axa*	2.7	2.7	1.0	2.7	2.3	3.0
x is taken from *ixi*	5.0	4.0	4.0	4.0	1.7	1.0

- transferable ω^x is taken from context $\left(\omega^i\left(t-1\right),\omega^x\left(t\right),\omega^i\left(t+1\right)\right)$ – 11% of records estimated <=2;
- transferable $\omega^x$is taken from context $\left(\omega^a\left(t-1\right),\omega^x\left(t\right),\omega^a\left(t+1\right)\right)$ – 44% of records estimated <=2.

2. Secondary words when transferable ω^x was taken from $(\omega^a(t-1)$, $\omega^x(t)$, $\omega^a(t+1))$ context received highest estimation. In this case total estimation (arithmetical mean of all 36 secondary words pronunciations estimations) is the lowest, i.e., secondary words pronunciations are the most similar to standard word pronunciations.
3. While analyzing two concrete ω^x allophones and when expert estimation is higher than 2, then it is necessary to synthesize different consonant ω^x allophones for high quality speech synthesis.

3.3. *Estimation of Automatic Speech Recognition Results*

Automatic speech recognition system's model was built for the experiments and speech recognition accuracy estimations were fixed. Below material describing experiments is given.

Speech recognition system was built using *Entropic HTK v.2.2* program package.

Corpus (speech database) for the experiments using $\left(\omega^m\left(t-1\right),\omega^x\left(t\right),\omega^n\left(t+1\right)\right)$, allophone combinations was created. Corpus is based on 180 phrases recorded by 10 speakers.

Corpus signals were digitized at 16KHz frequency. In signals preparation phase preemphasis (factor $= 0.97$) was applied. A 25ms Hamming analysis window function that was shifted with 10ms steps was used.

Mel-frequency cepstrum, first time-derivatives, acceleration and 0'th cepstral coefficients' vectors were prepared for recognition system analysis. These parameter vectors were selected taking into account previous experiment results. During previous experiments parameters influence for speech recognition resistance to noise were analyzed (Kandratavičius, 2000; Kandratavičius, 2001). The conclusions were drown that mel-frequency cepstrum coefficients are robust and give the best recognition accuracy.

Cepstral Mean Subtraction *CMS* method was used to increase prepared parameter vector robustness. This method was selected after noisy environment compensation methods

evaluation based on experimental application results (Kandratavičius, 1999). The main quality of this method – efficiency, i.e., using ordinary calculations high non-real time automatic speech recognition system's accuracy increase is received (Kandratavičius, 1999).

Prepared parameter (coeficient) vector (MFCC_0_D_A_Z) length is 39 elements.

Five states HMM models (left to right, context independent, the same to all allophones), described by one Gaussian mixture, were used.

Two types grammar was used for recognition system analysis:

1. when $(\omega^m(t-1), \omega^x(t), \omega^n(t+1))$ – 36 different allophones;
2. when $\left(\omega^m(t-1), \omega^{x'}(t), \omega^x(t), \omega^n(t+1)\right)$ – 6 different allophones.

Two different methods were used for training:

1. *Baum-Welch* algorithm (recognition results fixed after 7 training iterations were conducted);
2. *Viterbi* algorithm, analyzing isolated allophones, received by segmentation.

In order to maximize recognition system accuracy the same training allophones combinations for system testing were used.

Speech recognition accuracy evaluation (in percentage) is calculated by formula

$$Accuracy = \frac{N-K}{N} \cdot 100\%,$$

where N – general number of pronounced words in phrases used for testing; K – errors (incorrectly recognized allophones) number.

Results of experiments, received after analysis of recognition system results while using different grammars and different training scenarios are presented in Table 2.

Analyzing recognition results did recognition errors estimation.

During automatic speech recognition results' analysis, arithmetical means of all results were calculated. All results were transformed to 6-allophone grammar. Generalized experiments' results are given in Table 3 (where, e.g., *uxu* means ω^xtaken from $(\omega^u(t-1),\ \omega^x(t),\ \omega^u(t+1))$).

Next conclusions were drawn after experiments results' generalization.

1. Average automatic speech recognition's error rate – 22,7%.

Table 2

Recognition results

	Grammar	Parameter vector (length)	Training	
			Baum-Welch, 7 iterations Accuracy (%)	*Viterbi* Accuracy (%)
1.	$36\omega^x$	MFCC_0_D_A_Z (39)	77.4	76.9
2.	$6\omega^x$	MFCC_0_D_A_Z (39)	79.7	75.4

Table 3
Generalized confusion matrix

Recognized	*uxu*	*oxo*	*axa*	*exe*	*ėxė*	*ixi*
uxu	83.0%	13.8%	1.3%	0.9%	0.2%	0.7%
oxo	10.1%	85.6%	3.2%	0.2%	0.2%	0.6%
axa	1.8%	10.2%	80.8%	3.9%	2.5%	0.7%
exe	0.8%	2.7%	7.7%	70.3%	12.4%	6.0%
ėxė	1.6%	4.3%	5.0%	10.6%	62.3%	16.2%
ixi	0.9%	1.2%	2.6%	3.7%	9.5%	82.2%

2. 8,2% of errors are received while analyzing ω_k^x and ω_l^x allophones, when $|k-l| = 1$, 2,9% of errors are received when $|k-l| = 2$ and 1,2% of errors are received, when $|k-l| > 2$, where k and l are positional row indexes.
3. Allophones ω^x neighboring in positional row are in most cases incorrectly recognized. These allophones are the most resemblance.

The best recognition accuracy corresponds to the best expert evaluation mark (Table 3 and Table 1). This displays the correlation between expert evaluation and automatic speech recognition experiments' results.

4. Conclusions

Influence of different *ch* allophones for speech intelligibility and automatic speech recognition based on proposed basic frequency and positional raw was evaluated. Two types of experiments based on expert evaluation and on automatic speech recognition, were performed.

Experiment's results corroborated that *ch* allophones differ and depend from the context, i.e., from neighboring vowels. Results of automatic speech recognition corroborated, that neighboring in positional row *ch* allophones are the most resemblance. This corresponds to experts' evaluation conclusions, that the bigger the difference of context vowels' main frequency, the more *ch* allophones differs. Expert evaluation corroborated that different *ch* allophones have influence on speech intelligibility. Therefore for high quality speech synthesis different *ch* allophones must be synthesized.

References

Girdenis, A. (1981). *Phonology*. Mokslas, Vilnius (in Lithuanian).

Kandratavičius, I. (1999). Speech recognition in CTI applications. In *Informacinės technologijos'99*. Technologija, Kaunas. pp. 221–225 (in Lithuanian).

Kandratavičius, I. (2000). Voice signal flow sequencies and their resistance to noise. In *Informacinės technologijos'2000*. Technologija, Kaunas. pp. 145–150 (in Lithuanian).

Kandratavičius, I. (2001). Lithuanian speech corpora LTDIGITS: preparation and speech recognition results. *Informacinės technologijos'2001*. Technologija, Kaunas (in Lithuanian).

Pakerys, A. (1995). *Lithuanian Language Phonetics*. Žara, Vilnius (in Lithuanian).
Sonntag, G.P., T. Portele, F. Haas, J. Köhler (1999). Comparative evaluation of six German TTS systems. *Eurospeech'99*, 251–254.
Žintelytė M., I. Kandratavičius (1997). The problems and principles of speech processing. In *Informacinės technologijos'97*. Technologija, Kaunas. pp. 316–323 (in Lithuanian).
Žintelytė, M. (1999). Rewiev of speech synthesis methods. In *Informacinės technologijos'99*. Technologija, Kaunas. pp. 226–230 (in Lithuanian).
Žintelytė, M. (2000). Formation of synthesis element's set for Lithuanian language phonetic synthesizer. In *Informacinės technologijos'2000*. Technologija, Kaunas. pp. 156–161 (in Lithuanian).
Žintelytė, M. (2001). Phonetic specificity influence on high quality speech synthesis. In *Informacinės technologijos'2001*. Technologija, Kaunas (in Lithuanian).

M. Žintelytė has graduated from Kaunas University of Technology, Faculty of Informatics. She received her B.Sc. in 1993 and M.Sc. in 1995. Now she is a Ph.D. student at Kaunas University of Technology. Her research interest includes Lithuanian speech synthesis.

I. Kandratavičius has graduated from Kaunas University of Technology, Faculty of Informatics. He received his B.Sc. in 1993 and M.Sc. in 1995. Now he is a Ph.D. student at Kaunas University of Technology. His research interest includes Lithuanian speech recognition in noisy environment.

Sintezuojamo frikatyvo *ch* specifikos ir įtakos kalbos kokybei analizė

Marija ŽINTELYTĖ, Irmantas KANDRATAVIČIUS

Vienas iš sudėtingesnių sintezės uždavinių – nebalsingųjų frikatyvinių priebalsių sintezė. Atlikus priebalsio x fonetinę analizę buvo gauta pagrindinė išvada, kad priebalsį x įtakoja prieš ir po esantys fonetiniai elementai. Šiame straipsnyje aprašytų eksperimentų metu buvo siekiama įvertinti skirtingų x alofonų įtaką kalbos supratimui ir automatiniam kalbos atpažinimui.

Straipsnyje pateikiama automatizuotam kalbos sintezės klausimų sprendimui reikalinga formali sistema, nusakanti alofonų, o tuo pačiu fonemų, tarpusavio santykius galimose jų sekose natūralioje kalboje. Siekiant įvertinti skirtingų fonetinių vienetų ω^x tarpusavio panašumus buvo atlikti 2 skirtingų tipų eksperimentai: a) ekspertiniu vertinimu pagrįsta ω^x alofonų panašumo analizė; b) kalbos atpažinimo sistemos darbo rezultatų vertinimu pagrįsta fonetinių vienetų ω^x panašumo analizė.

Eksperimentų rezultatai patvirtino, kad skirtingų balsių kontekste esantys alofonai skiriasi savo akustinėmis savybėmis, įtakoja kalbos suprantamumą bei tai, jog sintezuojant aukštos kokybės lietuvių kalbą, būtina, priklausomai nuo konteksto, sintezuoti skirtingus x alofonus.